INDIVIDUAL DIFFERENCES AND THE EFFECTIVENESS OF VISUAL FEEDBACK ON REFLEXIVE BINDING IN L2 JAPANESE

A Dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Linguistics

By

Rebecca Raewyn Sachs, M.A.

Washington, DC
December 17, 2010
INDIVIDUAL DIFFERENCES AND THE EFFECTIVENESS OF VISUAL FEEDBACK ON REFLEXIVE BINDING IN L2 JAPANESE

Rebecca Raewyn Sachs, M.A.

Thesis Advisor: Alison Mackey, Ph.D.

ABSTRACT

Second language acquisition research into the effects of corrective feedback has investigated a variety of learning targets using a wide range of implicit and explicit feedback types (Li, 2010). To date, however, its linguistic focus has been limited to theoretically noticeable surface features (Carroll, 2001; Schmidt, 2001), and researchers have only recently begun to explore how individual differences might influence the effectiveness of particular feedback techniques (Robinson, 2005). This dissertation examines whether metalinguistic feedback presented in a visual format can help English-speaking learners improve their accuracy in interpreting the Japanese reflexive *zibun* (‘self’), whose binding behavior involves structural characteristics (e.g., c-command, subjecthood) not overtly represented in surface morphosyntax. Both positive and negative evidence may be necessary since, unlike English *himself*, Japanese *zibun* allows long-distance antecedents while disallowing non-subject antecedents (Thomas, 1995).

In a computer-assisted language learning experiment using a pre-/post-/delayed-post-test design, 80 university Japanese learners were randomly assigned into three treatment conditions: one informing them whether or not their interpretations were correct (Right/Wrong Feedback), one informing them of correctness and illustrating relevant structural relationships with tree diagrams (Trees Feedback), and one providing no feedback. Participants were also
administered a battery of cognitive tests, including the Modern Language Aptitude Test (Carroll & Sapon, 1959), the Visual Patterns Test (Della Sala et al., 1997), and a test of metalinguistic knowledge and sensitivity to ambiguity.

Results indicated clear evidence of improvement only in the Trees-Feedback group, and cognitive variables showed different relationships with test performance according to treatment condition: Visual memory, grammatical sensitivity, rote memory, and sensitivity to ambiguity predicted post-test scores in the Trees-Feedback condition; course levels and length of Japanese study predicted scores in the Right/Wrong-Feedback condition; and grammatical sensitivity, metalinguistic knowledge, and formal study of linguistics predicted scores in the No-Feedback condition.

Thus, visual feedback can facilitate the learning of a linguistic target which involves underlying structural relationships, but may work best for learners with particular aptitude profiles. Analyses of aptitude-treatment interactions hold great promise for advancing the field of SLA through producing insights about language learning processes and helping to optimize L2 instruction in relation to learners’ individual characteristics.
ACKNOWLEDGMENTS

This dissertation was truly a collaborative effort and would not have been possible without the assistance of many generous and incredibly supportive people.

The intellectual influence of my dissertation committee will, I hope, be obvious. Dr. Alison Mackey and Dr. Ronald Leow have been shaping my thinking about second language acquisition since the time I started at Georgetown, and echoes of their leading voices can be heard throughout these pages... although they might have been a bit more concise about things. I have also been extremely lucky to be able to work with Dr. Andrea Révész, a brilliant researcher whose meticulous approach to research design I can only hope to emulate. All three have been role models to me. Alison, Ron, and Andrea are true educators who genuinely care about their students and motivate us to shoot as high as possible. About Alison I can honestly say that I cannot imagine having a mentor more proactive about trying to open doors for me, challenging me to succeed and helping me to do so.

A major reason I ended up at Georgetown in the first place is thanks to Dr. Charlene Polio at Michigan State University. I went there with the goal of earning an M.A. in TESOL and, because of my experiences working with her, came out inspired to be a scholar. For the idea of a computer-mediated feedback experiment exploring the influence of a range of cognitive variables, I am indebted to Dr. Cristina Sanz and the formative experiences I had working with her on The Latin Project, as well as to the research of several students of hers, Ron’s, and Alison’s through the years. (Even a quick glance at the reference list will show how many previous Georgetown dissertations influenced this one.) A thought-provoking course taught by
Dr. Donna LARDIERE prompted the first research question of this dissertation, and her penetrating and insightful comments on materials design and papers leading up to this project have been invaluable. To Dr. Patrick Rebuschat I am grateful for an enlightening course on implicit learning, and for his suggestion that I include confidence judgments and source attributions on the pre- and post-tests. Dr. Yoshiko Mori’s office was always open for questions about Japanese linguistics, and her extensive knowledge in that area was instrumental in helping me to choose a linguistic target. Where statistical guidance is concerned, Dr. Rusan Chen has been an endless source of expert information, and I appreciate his advice not just on this project, but on several others as well. I have been honored to take and audit classes at the University of Maryland at College Park with Dr. Mike Long and Dr. Robert DeKeyser, whose sharp minds, strong opinions, and passion for the field of SLA are an inspiration to all of us. Robert not only sparked my interest in aptitude-treatment interactions, but went out of his way to provide me with an office on campus that I could use for data collection at UMD when the computer labs were occupied. Thanks are due also to Dr. Susanne Carroll, who, after meeting me just once at a conference, was kind enough to give me detailed, incisive, and encouraging feedback on an earlier paper that served as a foundation for this one.

Since I do not speak Japanese myself, it is quite literally true that this dissertation owes its existence to the help of several exceptionally generous and patient Japanese friends of mine. Mika Hama, quite possibly the most giving person I know, endured a near constant stream of questions about who *zibun* can and cannot refer to, as well as long stretches of nights spent converting all of the experimental materials into Japanese orthography and helping me to revise items based on pilot-testers’ reactions. When I was first trying to decide on a linguistic
target, Dr. Takako Egi provided translations, sentence judgments, and many helpful comments on particles. The infinitely patient Hirono Ishigami met with me every Wednesday for several weeks to create viable Japanese sentences containing the morphosyntactic characteristics I needed. Sakiko Kajino, with the assistance of Motoko Takada, then painstakingly checked every sentence to ensure that learners at an intermediate level or above might be expected to know the vocabulary and kanji. In the midst of busy schedules, Kumiko Akikawa, Shozo Aoki, Ayako Hama, Yosuke Miki, Kimi Nakatsukasa, and Wataru Suzuki pilot-tested all of the materials more than once and provided a great deal of helpful feedback. I owe them all my deepest gratitude.

For the working memory measures, I am grateful to Dr. Cristina Sanz, Dr. Catherine Stafford, and Dr. Harriet Wood Bowden, who let me use a listening-span test they had developed for The Latin Project with the assistance of Bill Garr of Georgetown’s Center for New Designs in Learning and Scholarship (CNDLS), and I am grateful to Dr. Jaemyung Goo, who let me adapt an operation-span test he had created. Dr. Heather Weger, whose knowledge of motivation research never ceases to impress me, designed the motivation questionnaires. Once data had been collected, Eunji Lee kindly served as a second rater on the test of sensitivity to ambiguity and knowledge of metalinguistic terms and helped me to revise their coding schemes.

For my meetings with participants, Mike Fekula, Facilities Coordinator at the University of Maryland, unfailingly put a smile on my face as the friendly and welcoming Keeper of Keys. Adam Hill, Executive Coordinator for Academic Technologies at The George Washington University, graciously helped me to find a computer lab there that I could use. At Georgetown, Dr. Peter Janssens and Susan Pennestri of CNDLS not only let me use their sophisticated
Language Learning Technology Center, but even arranged for me to be able to borrow laptops when my PowerPoints wouldn’t work on their beautiful Macs. At the beginning of the project, Peter and **Gorky Cruz** gave me much useful technological advice.

Research of this kind is not possible without participants, and I was incredibly fortunate that so many wonderful Japanese teachers welcomed me into their classrooms to recruit volunteers for the study. I am deeply grateful to **Dr. Yoshiko Mori**, **Motoko Omori**, **Kumi Sato**, and **Dr. Michael McCaskey** at Georgetown University; **Eiko Miura**, **Dr. Michele Mason**, **Etsuko Yamakita**, and **Makiko Inoue** at the University of Maryland at College Park; **Dr. Noriko Williams** at American University; and **Dr. Takaes Tsujioka**, **Mitsuyo Sato**, and **Rika Seya** at The George Washington University. All of them were very supportive, and I am sure that their positive energy about the project encouraged many students to participate. For making it all the way through the experiment, a million thanks are due to **Participants #1, 6, 7, 8, 9, 11, 14, 15, 16, 17, 20, 22, 24, 26, 27, 28, 29, 34, 35, 36, 37, 39, 40, 43, 45, 46, 47, 52, 54, 55, 56, 57, 59, 60, 61, 62, 63, 65, 66, 67, 68, 71, 72, 73, 74, 75, 76, 77, 79, 83, 87, 90, 93, 94, 95, 97, 98, 99, 100, 101, 103, 104, 105, 106, 107, 109, 110, 111, 113, 114, 117, 118, 120, 121, 122, 124, 126, 128, 129, 131, 132, 133, 135, and 136**, whom I’m not allowed to mention by name, but whose sticktoitiveness in the face of many, many other responsibilities I admire and greatly appreciate.

This project was supported by a generous Dissertation Grant from the journal **Language Learning**, and my studies at Georgetown were funded by the **Graduate School** and the **Department of Linguistics**, which provided me with tuition support, research and teaching assistantships, and conference travel grants for many years. **Manela Diez** and **Erin Esch** in the Linguistics Department have the ins and outs of every rule, guideline, and procedure imaginable.
at their fingertips, and I would especially like to thank Erin for going out of her way to keep me apprised of deadlines. I am grateful to Deanna Wormuth and Mandy Kama in the Department of English as a Foreign Language not only for gainful (and incredibly fulfilling) employment in an excellent program, but also for their personal support and encouragement.

Finally, I would like to acknowledge my friends and family, who are beyond value. I have already mentioned my angelic roommate Mika Hama. I would also like to thank Dr. Luis Cerezo, Dr. Missy Baralt, and the chupacabra, who helped to keep me laughing; Becca Duclos and Dr. Heather Weger, who are always there to ask and answer the Big Important Life Questions; Sean Vidanage, who taught me not to stress because all you can do is the best you can do; Bu, Msu, Gu, and Eddie, who stayed up with me during many late nights; Jacquelyn Kelly and her family, who have been like my family; and, of course, my own family, who have been my family: my grandparents, aunts, uncles, and cousins; Dr. Ben Sachs, the best, smartest, and most hilarious little brother in the world (want a raisin?); Diane Patricia Sirois, the epitome of unconditional love and support, who is still my ideal of exactly who I want to be when I grow up; and Dr. Paul F. Sachs, who can do absolutely anything he sets his mind to and has always tried to convince me and Ben that we can, too. To Justin Kelly, I won’t say anything sappy, but I will tell you this: I know that it’s demanding to defeat those evil machines, but I know you can beat them. Then we can seek sound shelter and classify water damage as art—we don’t need no walkie-talkies; we can make it if we run!
This dissertation is dedicated to my parents.
TABLE OF CONTENTS

CHAPTER 1: FEEDBACK .............................................................................................................. 1
The effectiveness of feedback for second language learning ............................................... 1
  Reconciling theory and empirical research ....................................................................... 1
  Feedback and the weak interface ..................................................................................... 8
Empirical studies comparing explicit and implicit feedback ............................................... 14
  Advantages found for explicit feedback .......................................................................... 17
  No advantages found for explicit feedback ..................................................................... 30
Learner awareness of information provided in feedback ................................................... 37
  Carroll’s ‘awareness constraint on negative evidence’ ................................................... 41

CHAPTER 2: APTITUDE FOR LANGUAGE LEARNING .............................................................. 53
(Re)conceptualizing aptitude .............................................................................................. 53
  The Modern Language Aptitude Test .............................................................................. 54
  Updating and situating the concept of aptitude ............................................................. 57
  Working memory ......................................................................................................... 59
  Familiarity with metalinguistic concepts and terminology ............................................ 63
Relevance of language aptitude across learning contexts .................................................. 73
Aptitude-treatment interactions ......................................................................................... 75
  Classroom-based studies of aptitude-treatment interactions ........................................ 78
  Laboratory-based studies of aptitude-treatment interactions ....................................... 87

CHAPTER 3: REFLEXIVES IN A SECOND LANGUAGE .............................................................. 110
Theories of reflexives employed in empirical studies of their acquisition ......................... 110
  Manzini and Wexler’s parameterized approach ........................................................... 112
  Progovac’s Relativized SUBJECT approach ............................................................... 115
  LF-movement approaches ............................................................................................. 118
Previous empirical research on the second language acquisition of reflexives ................ 123
  Results interpreted as evidence of learners following universal principles ................. 124
  Results interpreted as evidence against the Subset Principle ....................................... 126
Long-distance binding and subject orientation ............................................................. 130
Results interpreted as evidence for first-language influence ....................................... 134
Results interpreted as evidence against first-language influence ................................ 140
No apparent L1 influence on L2 binding patterns ..................................................... 140
The finite/nonfinite asymmetry ................................................................................. 142
Long-distance binding and morphology ........................................................................ 149
Morphological triggers? .................................................................................................... 159
Attempts at instructional intervention ........................................................................ 164

CHAPTER 4: METHODS .......................................................................................................... 173
Research questions ........................................................................................................... 173
Participants........................................................................................................................ 175
Linguistic target ................................................................................................................. 180
Overview of design .............................................................................................................. 187
Recruitment of volunteers and the informed consent process ....................................... 191
Design of the tests and treatments ................................................................................... 194
  Truth-value judgment task ............................................................................................ 194
  Treatment conditions .................................................................................................... 201
Sentence types used in the tests and treatments ......................................................... 209
Counterbalancing of test forms ..................................................................................... 215
Piloting of the testing and treatment items .................................................................. 217
Test scoring and reliability ............................................................................................. 221
Concurrent verbal protocols ........................................................................................... 225
Individual-difference measures ..................................................................................... 229
  Background questionnaire ............................................................................................ 229
  Test of sensitivity to ambiguity and knowledge of metalinguistic terminology .......... 230
    Ambiguity section ........................................................................................................ 230
    Metalinguistic terminology section ........................................................................... 235
  The Modern Language Aptitude Test ........................................................................ 239
  The Visual Patterns Test ................................................................................................. 242
Summary of variables ....................................................................................................... 246
<table>
<thead>
<tr>
<th>Appendix</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2</td>
<td>Instructions for the treatments</td>
<td>375</td>
</tr>
<tr>
<td>E3</td>
<td>Instructions for thinking aloud</td>
<td>378</td>
</tr>
<tr>
<td>F</td>
<td>Zibun sentences and interpretations</td>
<td>379</td>
</tr>
<tr>
<td>F1</td>
<td>List of sentences used on the tests of zibun interpretation</td>
<td>379</td>
</tr>
<tr>
<td>F2</td>
<td>List of sentences used in the treatments</td>
<td>385</td>
</tr>
<tr>
<td>F3</td>
<td>Characteristics of zibun</td>
<td>390</td>
</tr>
<tr>
<td>G</td>
<td>Motivation and reflection questionnaires</td>
<td>397</td>
</tr>
<tr>
<td>G1</td>
<td>Motivation questionnaire—Preactional</td>
<td>398</td>
</tr>
<tr>
<td>G2</td>
<td>Motivation questionnaire—Actional #1</td>
<td>400</td>
</tr>
<tr>
<td>G3</td>
<td>Motivation questionnaire—Actional #2</td>
<td>401</td>
</tr>
<tr>
<td>G4</td>
<td>Motivation questionnaire—Post-actional</td>
<td>402</td>
</tr>
<tr>
<td>G5</td>
<td>Reflection questionnaire</td>
<td>406</td>
</tr>
<tr>
<td>H</td>
<td>Tests of working memory capacity</td>
<td>407</td>
</tr>
<tr>
<td>H1</td>
<td>Listening-span test answer sheet</td>
<td>409</td>
</tr>
<tr>
<td>H2</td>
<td>Operation-span test answer sheet</td>
<td>410</td>
</tr>
<tr>
<td>H3</td>
<td>Instructions for the operation-span test</td>
<td>412</td>
</tr>
<tr>
<td>I</td>
<td>Test of sensitivity to ambiguity and familiarity with metalinguistic terms</td>
<td>414</td>
</tr>
<tr>
<td>J</td>
<td>Visual Patterns Test sample answer sheet</td>
<td>418</td>
</tr>
<tr>
<td>K</td>
<td>Background questionnaire</td>
<td>419</td>
</tr>
<tr>
<td>L</td>
<td>Reactivity of thinking aloud</td>
<td>420</td>
</tr>
<tr>
<td>M</td>
<td>Latency of thinking aloud</td>
<td>425</td>
</tr>
<tr>
<td>N</td>
<td>Parallel coordinate plots showing test performance of each learner</td>
<td>432</td>
</tr>
<tr>
<td>O</td>
<td>Boxplots and histograms for individual-difference measures</td>
<td>435</td>
</tr>
<tr>
<td>O1</td>
<td>Visual Patterns Test</td>
<td>435</td>
</tr>
<tr>
<td>O2</td>
<td>Modern Language Aptitude Test</td>
<td>436</td>
</tr>
<tr>
<td>O3</td>
<td>Test of sensitivity to ambiguity and familiarity with metalinguistic terminology</td>
<td>437</td>
</tr>
<tr>
<td>O4</td>
<td>Test of sensitivity to ambiguity in English</td>
<td>438</td>
</tr>
<tr>
<td>O5</td>
<td>Test of familiarity with metalinguistic terminology</td>
<td>439</td>
</tr>
<tr>
<td>O6</td>
<td>Years of Japanese study</td>
<td>440</td>
</tr>
<tr>
<td>O7</td>
<td>Japanese course levels</td>
<td>442</td>
</tr>
</tbody>
</table>
### List of Figures

**Figure 2.1.** Example of an artificial grammar used in experiments by Reber .......... 93

**Figure 4.1.** C-command requirement on antecedents of reflexives .................. 182

**Figure 4.2.** Overview of the research procedure ........................................ 190

**Figure 4.3.** Example of two items associated with a causative sentence ........... 199

**Figure 4.4.** Test slide asking for a confidence judgment and source attribution .... 200

**Figure 4.5.** Examples of encouragement slides in the No-Feedback condition ....... 204

**Figure 4.6.** Examples of the feedback provided in the Right/Wrong condition ....... 205

**Figure 4.7.** Diagrams provided in the Trees condition for the causative sentence shown in Figure 4.3 ................................................................. 206

**Figure 4.8.** Introduction to tree diagrams .................................................. 207

**Figure 4.9.** Two sample items from Della Sala et al.’s (1997) Visual Patterns Test .... 243

**Figure 5.1.** Boxplots of test scores across treatment conditions ....................... 250

**Figure 5.2.** Comparison of the treatment conditions with regard to mean test scores ................................................................. 258

**Figure 5.3.** Scatterplots of relationships between VPT and pre-/post-test scores, with a separate regression line for each group ...................... 290

**Figure 5.4.** Scatterplots of relationships between MLAT IV and pre-/post-test scores, with a separate regression line for each group ................. 291

**Figure 5.5.** Scatterplots of relationships between MLAT V and pre-/post-test scores, with reference lines at 80% on each measure ......................... 292

**Figure 6.1.** Comparison of treatment groups’ test scores, including only Georgetown and the University of Maryland ................................. 308

**Figure 6.2.** Scatterplots of relationships between MLAT IV and pre-/post-test scores, highlighting the Right/Wrong-Feedback condition .......... 328

**Figure 6.3.** Scatterplots of relationships between years of study and pre-/post-test scores, highlighting the Trees-Feedback condition ................. 333

**Figure 6.4.** Scatterplots of relationships between course levels and pre-/post-test scores, highlighting the Trees-Feedback condition ........................ 334
LIST OF TABLES

**Table 1.1.** Linguistic targets in studies on the effectiveness of feedback ................. 49

**Table 2.1.** Aptitude components and awareness reports associated with learning easy and hard rules across conditions in Robinson (1997b) ................................. 91

**Table 2.2.** Tests showing statistically significant correlations between IDs and L2 performance for each rule type in Robinson (2002) .............................................. 100

**Table 2.3.** Tests showing statistically significant correlations between cognitive capacities and L3 performance in Lado (2008) ..................................................... 106

**Table 4.1.** Number of participants at each course level ........................................ 177

**Table 4.2.** Overview of test- and treatment-item characteristics ............................ 214

**Table 4.3.** Cronbach’s $\alpha$ for each test version across experimental groups .......... 222

**Table 5.1.** Goodness-of-fit tests to check for normality of distributions on the pre-, post-, and delayed post-tests across groups .............................................. 251

**Table 5.2.** Descriptive statistics of pre-, post-, and delayed post-test scores across groups ........................................................................................................... 252

**Table 5.3.** Tests of homogeneity and sphericity assumptions for Research Question 1 .................................................................................................................. 254

**Table 5.4.** Separate RM ANOVAs testing for improvement over time in each treatment group ........................................................................................................... 257

**Table 5.5.** Results of RM ANOVA comparing test scores in the 3 treatment groups over time ........................................................................................................... 261

**Table 5.6.** Descriptive statistics and tests of homogeneity for each ID measure ...... 264

**Table 5.7.** Skewness and kurtosis ratios for each ID, both overall and by treatment group .............................................................................................................. 265

**Table 5.8a.** Bivariate Spearman correlations between ID measures and test scores, and partial correlations between IDs and post-test scores controlling for pre-test scores: All participants .................................................. 277

**Table 5.8b.** Bivariate Spearman correlations between ID measures and delayed test scores, and partial correlations between IDs and delayed test scores controlling for pre-test scores: All participants .................................................. 278

**Table 5.9a.** Bivariate Spearman correlations between ID measures and test scores, and partial correlations between IDs and post-test scores controlling for pre-test scores: No-Feedback condition ........................................ 279
Table 5.9b. Bivariate Spearman correlations between ID measures and delayed test scores, and partial correlations between IDs and delayed test scores controlling for pre-test scores: No-Feedback condition ............... 280

Table 5.10a. Bivariate Spearman correlations between ID measures and test scores, and partial correlations between IDs and post-test scores controlling for pre-test scores: Right/Wrong condition ........................................ 281

Table 5.10b. Bivariate Spearman correlations between ID measures and delayed test scores, and partial correlations between IDs and delayed test scores controlling for pre-test scores: Right/Wrong condition .................... 282

Table 5.11a. Bivariate Spearman correlations between ID measures and test scores, and partial correlations between IDs and post-test scores controlling for pre-test scores: Trees condition ......................................... 283

Table 5.11b. Bivariate Spearman correlations between ID measures and delayed test scores, and partial correlations between IDs and delayed test scores controlling for pre-test scores: Trees condition ................................. 284

Table 5.12. Intercorrelations among individual-difference variables ..................... 297

Table 5.13. Multiple regression model predicting post-test scores in the Right/Wrong group ........................................................................................................ 299

Table 5.14. Multiple regression model predicting post-test scores in the Trees group ............................................................................................................. 300

Table 5.15. Another multiple regression model predicting post-test scores in the Trees group .................................................................................................. 300

Table 6.1. Descriptive statistics of test scores across conditions, including only GU and UMD ........................................................................................................... 309

Table 6.2. Statistically significant correlations between IDs and test scores (all participants) ......................................................................................................................... 318

Table 6.3. Statistically significant correlations between IDs and test scores (by treatment group) ........................................................................................................... 319
CHAPTER 1

FEEDBACK

The effectiveness of feedback for second language learning

Reconciling theory and empirical research

The issues of whether and how corrective feedback can play a causal role in creating new grammatical knowledge in a second language (L2) generate a good deal of interest even when they are not the subject of major controversies. L2 practitioners are keen to understand how feedback’s potential might be harnessed most efficiently for instructional purposes, while L2 theorists continue to deliberate what its effectiveness (or a lack thereof) says about the processes of adult second language acquisition (SLA). Historically, debates over L2 feedback have reflected important disagreements over which types of linguistic input are usable for SLA and which are not. Some researchers (e.g., Gregg, 1996; Krashen, 1981, 1985; Schwartz, 1993, 1999; Truscott, 1996, 1999) have argued that only ‘comprehensible input’ (or ‘primary linguistic data’ or ‘positive evidence’) can lead to language acquisition, their view being that although metalinguistic information and corrections (or ‘negative data’) might contribute to the conscious control or monitoring of language behavior, they cannot affect learners’ competence or spontaneous language use. Meanwhile, others (e.g., N. Ellis, 2005; Long, 1996; White, 1991) have contended that corrective feedback can be facilitative of L2 development, emphasizing that negative evidence might be necessary for learners to gain knowledge of certain contrasts between the first language (L1) and the L2.
Theoretical disagreements concerning types of evidence are not currently at the forefront of discussions in the L2 feedback literature, but teachers and researchers alike still have many reasons to debate whether, when, and how L2 learners’ errors should be corrected. At least four factors give cause for optimism that a greater understanding can be reached: (1) a proliferation of ever more rigorous empirical research in the area of L2 feedback; (2) several recent meta-analyses indicating that feedback is beneficial (Li, 2010; Lyster & Saito, 2010; Mackey & Goo, 2007; Russell & Spada, 2006); (3) greater recognition that “there is more than one route to L2 development” (Mackey, 2007, p. 24; see also Skehan, 1998); and (4) more nuanced perspectives on relationships between explicit and implicit knowledge, informed by cognitive psychology and neuroscience (e.g., N. Ellis, 2005).

Among the outcomes of the past decade’s rapid increase in empirical studies on the effectiveness of L2 feedback (demonstrated by Li, 2010, p. 310) and the fact that work in this area is adhering to higher research standards (according to Russell & Spada, 2006, p. 156), there has been “growing evidence that [corrective feedback] can be helpful for L2 learning” (p. 135), with “substantial” and “durable” effects (p. 152). Many recent studies have shown feedback to have a positive influence on L2 development (e.g., Ammar, 2008; Ammar & Spada, 2006; Ayoun, 2001; Carroll & Swain, 1993; Doughty & Varela, 1998; Egi, 2007b; R. Ellis, 2007; R. Ellis, Loewen, & Erlam, 2006; Han, 2002; Ishida, 2004; Iwashita, 2003; Leeman, 2003; Loewen & Nabei, 2007; Loewen & Philp, 2006; Long, Inagaki, & Ortega, 1998; Lyster, 2004; Lyster & Izquierdo, 2009; Mackey & Philp, 1998; McDonough, 2005, 2007; McDonough & Mackey, 2006; Muranoi, 2000; Nagata, 1993, 1995; Nagata & Swisher, 1995; Révész, 2009; Sachs & Suh, 2007; Sagarra, 2007; Sheen, 2007), and many of them have employed tests designed to measure
implicit L2 knowledge or communicative outcome measures encouraging spontaneous language production, calling into question the idea that feedback cannot affect (at least relatively) unmonitored language use. The meta-analyses mentioned above—Russell and Spada’s (2006) examination of research concerning grammar-oriented feedback on learners’ oral and written L2 production; Mackey and Goo’s (2007) examination of research investigating negotiated L2 interaction; Li’s (2010) examination of feedback research, which also included computer-delivered feedback; and Lyster and Saito’s (2010) examination of corrective feedback research in classroom settings—have found medium to large effect sizes when comparing feedback conditions against conditions without (or with less) feedback, including on delayed post-tests.¹

More than a decade ago, Lyster, Lightbown, and Spada (1999) argued that there was “increasing evidence that feedback on error can be effective” and that what was needed at the time was systematic research into the influence of feedback type, instructional context, and learner characteristics (p. 464). Li’s (2010) recommendation, based on more recent meta-analytic results, is that since “the effect of corrective feedback has been established, researchers should embark on the mission of investigating the factors constraining its effectiveness” (p. 349). Current investigations are, in fact, quite often focusing on identifying which characteristics and contexts of feedback are most strongly related to its uses and effectiveness for SLA, exploring, for example, the communicative contexts of learners’ nontargetlike utterances (Oliver & Mackey, 2003); the characteristics of communicative tasks, including their cognitive demands and learners’ familiarity with their content (Révész, 2009; ¹ Spada and Tomita’s (2010) meta-analysis of the effectiveness of L2 instruction more generally has also shown that explicit instruction has a positive impact on both controlled and spontaneous language use.
Révész & Han, 2006); the length and prosodic features of feedback moves, such as the use of emphatic stress and intonation to highlight corrections (Loewen & Philp, 2006); the number of changes a piece of feedback makes in relation to the learner’s original utterance (Philp, 2003); the linguistic focus of the correction (e.g., morphosyntax vs. lexicon) (Havranek, 2002; Jeon, 2007); whether the feedback provides or elicits more targetlike forms (Ammar, 2008; Ammar & Spada, 2006; Lyster, 2004; Lyster & Izquierdo, 2009; Nassaji, 2009); the amount of metalinguistic detail it contains (Nagata, 1993, 1995; Nagata & Swisher, 1995; Rosa & Leow, 2004b; Sanz & Morgan-Short, 2004); learners’ responses to feedback (e.g., McDonough & Mackey, 2006; Révész, Sachs, & Mackey, to appear); their proficiency or developmental readiness (Iwashita, 2003; Mackey & Philp, 1998); cognitive factors and other individual differences (e.g., Mackey, Philp, Fujii, Egi, & Tatsumi, 2002; Mackey, Adams, Stafford, & Winke, 2010; Révész, to appear (a), to appear (b); Sagarra, 2007; Sheen, 2007; Trofimovich, Ammar, & Gatbonton, 2007); learners’ interpretations of feedback (e.g., Carpenter, Jeon, MacGregor, & Mackey, 2006; Gass & Lewis, 2007; Egi, 2007a, 2007b, 2010; Kim & Han, 2007; Mackey, Gass, & McDonough, 2000); and orientations to feedback in different classroom contexts (Lyster & Mori, 2006; Sheen, 2004).

Li’s (2010) meta-analysis suggests that the greatest effects for feedback may tend to be found when experiments are carried out with discrete-item practice sessions, in laboratory-based settings, with native speakers providing the corrections. However, communicative, classroom-based studies of L2 feedback have shown substantial effect sizes as well, and Lyster and Saito’s (2010) meta-analysis of feedback in classroom settings has revealed that “students in L2 classrooms appear to exhibit improvement more clearly in free constructed-response
measures than in metalinguistic tasks” (p. 285), a finding they explain (speculatively) with reference to transfer-appropriate processing (p. 292). In other words, it seems likely that many of the abovementioned factors interact with each other in intricate ways. According to Leeman (2007), there is “still no definitive answer to the question of how feedback promotes L2 development nor of what is the specific contribution of the various components of feedback” (p. 128). Some might contend that, considering the wide range of potentially relevant contextual factors, individual learner differences, and other variables involved, there never will be cut-and-dried answers to these questions. Nonetheless, given the research developments mentioned above and discussed further below, debates can now be less impeded by the formerly oft-repeated verdict that negative evidence does not directly affect L2 competence. As Leeman (2007) points out, “acknowledging the value of feedback does not require the adoption of any one theoretical perspective…. Indeed, [a wide] gamut of theoretical perspectives... is consistent with the results from empirical feedback research” (p. 129).

In relation to some approaches to SLA, the connections are straightforward. In MacWhinney’s (2001, 2011) Competition Model, for instance, SLA is driven by learners’ evolving sense for which cues to form-meaning connections are the strongest and most reliable. Learners are “heavily influenced by transfer” from the L1, features of which may be entrenched in ways that limit the detectability of crucial L2 cues (2011, p. 4). In this context, advance organizers based on metalinguistic information and/or feedback can help learners to focus their attention in helpful ways. Negative feedback can indicate that a form-meaning mapping is not valid, thereby helping to “tune cue weights” (p. 18), and meanwhile, albeit perhaps very incrementally, “positive evidence for any one form (or meaning) effectively constitutes negative

In skill acquisition theory (e.g., DeKeyser, 2007), different types or aspects of feedback may play different roles and be more or less helpful according to learners’ stages in the processes of proceduralization and automatization. Initially, feedback might facilitate the attainment of declarative knowledge; thereafter, it might help learners to avoid proceduralizing or automatizing nontargetlike production by indicating where greater attention might be needed. Errors arise from different sources (e.g., a lack of competence, lapses in attention, misguided declarative knowledge, overgeneralizations), and the relevance, effectiveness, and efficiency of various kinds of feedback will sometimes depend on these factors (Leeman, 2007).

As far as generative SLA theory is concerned, recognizing (as generative theorists do—see, e.g., Beck, Schwartz, & Eubank, 1995) that there are distinctions among negative evidence, negative data, and negative feedback (‘data’ being information; ‘feedback’ being a discourse move which can provide negative data, positive data, or both; and ‘evidence’ referring to the function of the information), it may be possible to achieve some degree of reconciliation between theoretical claims and empirical findings which might otherwise seem incompatible. Beck et al. (1995) state that “different types of data are only potential evidence for grammar building. While all evidence for constructing a grammar comes from data, not all data turn out to be evidence” (p. 177). As possible sources of data, they identify (1) contextualized utterances in ambient language (i.e., positive data), (2) implicit or explicit information about the impossibility of a form or utterance (i.e., negative data), and (3) descriptive information about the L2 (i.e., explicit data). Of these, they argue that only the first (positive data) can be a source
of evidence for grammar building; “negative data and explicit data cannot be used in the creation of interlanguage (or L1) grammars” (Schwartz, 1999 pp. 650-651).

Now, descriptive and experimental empirical studies conducted with different priorities and assumptions have indicated that when negative feedback is provided, learners at least sometimes perceive it as corrective and make use of it (as measured by immediate modified output or later grammatical accuracy, for example). These findings do not necessarily mean that the feedback has imparted negative evidence, however, and SLA researchers working in a generativist paradigm may therefore be able to maintain the assertion that negative data do not directly affect L2 competence: Perhaps the beneficial effects on L2 development have been due in some cases to co-occurring salient positive evidence (Leeman, 2003); to the learners’ own modified output in response to feedback, which then served as positive self-input (Sharwood Smith, 1981, 1991); to the fact that the learners were ‘pushed’ to expand their linguistic repertoires (Swain, 1995, 2005); etc. In certain cases, though, the learning which has been demonstrated has seemed to suggest that problems of blame assignment, often mentioned as a major obstacle to the effectiveness of feedback (e.g., Carroll, 2001; see Pinker, 1989), can either be surmounted or circumvented. Moreover, it is worth considering practical efficiency for L2 pedagogy in addition to being concerned with theoretical issues (Long, 2007).

These latter points, along with the growing empirical evidence for facilitative effects of feedback on communicatively oriented measures of L2 development reviewed above, seem worthy of greater acknowledgment than they are perhaps given in generative SLA research, where a common assumption may be that the question of negative feedback’s effectiveness
has already been settled or is simply no longer of interest. Generative SLA researchers do not tend to conduct experimental (pre-test/post-test) studies of L2 feedback and may not always be very familiar with the results of research in that area. The converse, of course, may often be true regarding L2 feedback researchers’ familiarity with formal linguistic analyses. Whereas generative researchers have conducted extensive research into the L2 acquisition of reflexive binding, for example (as reviewed in Chapter 3), L2 feedback researchers have not focused on reflexives as a linguistic target and may have only vague knowledge of theories concerning their acquisition. In other words, the two groups might stand to learn quite a lot from each other, and the field of SLA as a whole might benefit from their collaboration.

Feedback and the weak interface

Krashen’s (1981, 1985) distinction between learning and acquisition, and Schwartz’s (1993) distinction between ‘learned linguistic knowledge’ and ‘linguistic competence’, remain astute observations and are acknowledged as such by researchers operating under diverse theoretical frameworks (see R. Ellis, 2004). “[I]n psychological vernacular,” as N. Ellis (2005) puts it, “explicit and implicit knowledge are distinct and dissociated; they involve different types of representation and are substantiated in separate parts of the brain…. [E]xplicit knowledge does not become implicit knowledge, nor can it be converted to it” (p. 307). “[A]ny cognitive psychology or neurobiology text will affirm” that they are functionally and anatomically separate (p. 340). Nonetheless, N. Ellis continues, “conscious and unconscious processes are dynamically involved together in every cognitive task and in every learning episode” (p. 340);

---

2 The impression expressed here is based on a general sense of the topics of papers which tend to be presented at conferences and published in journals focusing on generative SLA.
explicit knowledge can shape language processing in the sense that different aspects of the primary linguistic data may receive more focal attention when it is present. As a simple demonstration, N. Ellis (2005) entreats his audience to read the following (p. 325):

[pay] bill [speed?] sprinted [when?] crawls
[duck] bill [when?] sprinted [speed?] crawls
[had to] bill [rhyme?] sprinted [person?] crawls

He explains that a person’s reading of the second item in each pair is likely to differ from line to line because ‘[when?]’ focuses on tense, ‘[speed?]’ on meaning, and so on. Explicit information suggests a context and thereby guides or constrains interpretation, which leads to “different tallying and different tuning” by the implicit system (p. 325).

To tie this in with L2 feedback: In adult second language acquisition, L1 influence can sometimes impede the sorts of language processing that would result in more targetlike L2 knowledge. Automatization is “the acme of L1 learning” but “the bane of SLA” (N. Ellis, 2005, p. 327), and a “sad irony” is that simply receiving more positive input or producing more output can be counterproductive; L2 learners risk “[digging] themselves ever deeper” (p. 326). This is therefore a situation in which explicit metalinguistic information—provided in feedback, for example—might be helpful. Schmidt (2001) points out that “since many features of L2 input are likely to be infrequent, non-salient, and communicatively redundant, intentionally focused attention may be a practical (though not theoretical) necessity for successful language learning” (p. 23). If immediate corrections can change the focus of learners’ attention and induce them to modify form-meaning mappings upon realizing that their previous ones were inaccurate, the corrections may affect what serves as intake.
On an opposing side of the argument, Schwartz (1999) has used the Müller-Lyre illusion as an analogy to illustrate why “encyclopedic knowledge cannot shape competence” (p. 650). It is an apt example, and Schwartz makes an excellent point, well-taken as far as it goes:

The reason Learned knowledge cannot turn into Acquired knowledge... is for exactly the same reason the Müller-Lyre line illusion... remains an illusion [i.e., does not allow perception of the fact that the horizontal lines are equal in length] despite our knowing and believing that the two lines are equal in length. This reason, in two words, is information encapsulation... [which] prevents information outside a module...—in this case, encyclopedic knowledge that happens to be about language—from entering the encapsulated module—in this case, the language module. Thus, what appears to be a potential source of data for the language acquisition device... in actuality is not (p. 650).

Schwartz’s (1999) references to informational encapsulation originate in Fodor’s (1983) theory of modularity. Regarding feedback, she states that the question of whether correction is helpful should be reformulated to ask, “Can the knowledge that results from learning through correction serve as input to the language module?” (p. 650; Schwartz, 1993, p. 157). She answers in the negative, specifying that she is not claiming that explicit information cannot give rise to any knowledge, but rather simply that it cannot give rise to the sort of knowledge developed in L1 acquisition.
Considering the arguments above, however, it seems fairly clear that even researchers taking quite different angles on the debate would agree with Schwartz that the answer to her question is no. Non-generativists would not try to argue that “facts about language (or line length)... [can] be integrated into the language (or vision) module” (p. 650) or that explicit knowledge can be ‘transformed’ directly into implicit knowledge. Granted, some non-generativists (as well as generativists) have argued that adult L2 learners can develop the ability to use metalinguistic rules more quickly, but Paradis (2004) contends that this still involves controlled processing, simply sped-up (cited by N. Ellis, 2005, p. 333). Even DeKeyser (1997, 2003), who argues that proceduralized explicit knowledge is qualitatively different from declarative knowledge, suggests only that the former might be considered “functionally equivalent” to implicit knowledge. Summarizing points made by Hulstijn (2002) and N. Ellis (1994), R. Ellis (2004) states, “it is not the rules themselves that become implicit, but rather the sequences of language that the rules are used to construct” (p. 238). More interesting and productive questions, therefore, would seem to be, “How can the knowledge that results from learning through correction shape input to the language module?” and “How can this be harnessed for more efficient SLA?”

The notion of a so-called ‘weak interface’, in which explicit information that “come[s] to consciousness at the appropriate moment” (N. Ellis, 2005, p. 331) might guide learners’ processing of primary linguistic data and thereby influence acquisition indirectly, has been around for years (e.g., N. Ellis, 1994; R. Ellis, 1993; Schmidt, 1994). Certainly, this must have limits, and, despite agreeing with the fundamental idea, Paradis (2009) has argued at length, vehemently, that explicit guidance of this sort should not be called an interface, technically
speaking. A problem with strong no-interface positions, however, is that they do not seem to acknowledge sufficiently the types of influences which explicit knowledge can exert on the parallel development of implicit knowledge.

Returning for a moment to the Müller-Lyre illusion: It is true that individuals with the gift of eyesight cannot help but interpret the diagonal lines in the picture automatically as perspectival cues. It seems reasonable to assume that no amount of training would enable us to stop doing so. However, can we necessarily assume that the analogy holds for all areas of an L2 where negative evidence might theoretically be necessary for learning? The reason the visual illusion works may be because, in human vision, perspective functions in such a way that diagonal lines converging toward each other from opposite sides of a display produce the sense that whatever appears between their converging endpoints is farther away. This, combined with our experience that objects at a distance are actually larger than they appear to be, may automatically produce the perception that whatever is between the lines’ converging endpoints is larger than it appears. (The opposite is true for whatever appears between diagonal lines oriented in the opposite way.) In the Müller-Lyre illusion, this means that the horizontal line on the left looks closer to us, while the one on the right looks farther away and is automatically interpreted as being larger than it appears. In the real world, this innate universal (activated in all humans with early exposure to visual stimuli) might help all of us, as humans, to evade tigers most of the time. In language, innate universals (theoretically activated in all humans with early exposure to linguistic stimuli) might help all of us, as humans, to avoid subjacency violations most of the time, for whatever reason. Nothing lets vision work otherwise, and nothing apparently lets language work otherwise.
Nevertheless, there are also aspects of language that do differ cross-linguistically and for which there are not necessarily such strong predispositions against the possibility of modifying their processing and production with the help of explicit information. A person may have grown up raising main verbs over adverbs; using word order instead of case morphology as a primary cue to semantic roles; interpreting direct objects, but not long-distance subjects, as possible antecedents for reflexives; perceiving both [p] and [pʰ] as /p/; and so on; and it may be exceedingly difficult to change these developed biases, perhaps particularly in cases where negative evidence would seem to be required. However, they are not universally hardwired, and the observation that L2 learners “can show explicit knowledge about the target language while simultaneously failing to put this knowledge to use” (Schwartz, 1999, p. 649) does not necessarily mean that explicit knowledge (e.g., about L2-specific attentional strategies) can never have an influence on SLA in the long run if repeatedly put to use in helpful ways during processing (e.g., VanPatten & Cadierno, 1993; VanPatten & Oikhenon, 1996).

In sum, while it is, indeed, of theoretical importance to know whether negative evidence can affect learners’ L2 competence directly, it is also crucial to recognize the robust empirical evidence demonstrating that feedback on learners’ errors can affect (at least) L2 performance across several areas of language, with enduring benefits. Reconceptualizations of the so-called ‘interface’ between explicit and implicit knowledge might help to reconcile generative theorizing with promising experimental results from the large body of L2 feedback research. It is justifiable for Universal Grammar (UG) theorists to continue to assert that learned linguistic knowledge cannot be integrated directly into the language module given the incompatibility of the formats in which the information is couched. In certain senses, that is widely agreed upon
by non-generativists as well (connectionists, cognitive psychologists, etc.). However, if negative data, in the form of explicit / declarative / metalinguistic information, can affect competence indirectly through directing attention or supplying positive input that implicit / procedural / acquisitional processes can work on in parallel, then generativists’ rigorous cross-linguistic analyses and predictions about where negative evidence should theoretically be necessary could be productively put to use in designing studies with experimental treatments, pre-test/post-test designs, and ultimately, perhaps, implications for pedagogy. Even maintaining their ideas about informational encapsulation and modularity, generative researchers should be able to help learners to develop L2 competence more efficiently with the (indirect) assistance of metalinguistic information that reaches levels of accuracy and completeness far beyond what is often provided via pedagogical rules of thumb (and in non-generative L2 feedback research).

Empirical studies comparing explicit and implicit feedback

Back, then, to the “mission of investigating the factors constraining [feedback’s] effectiveness” (Li, 2010, p. 349). As mentioned above, L2 feedback researchers do not have definitive answers to the questions of which types or components of feedback are most effective in which areas of language, at which stages of learning, for learners with which characteristics, and so on. Concerning comparisons of ‘explicit’ versus ‘implicit’ feedback, a lack of statistically significant differences in meta-analyses of feedback research corroborates appropriately cautious statements by authors of primary studies to the effect that while several experiments have indicated that explicit feedback is more beneficial than implicit feedback (e.g.,
Bowles, 2005; Carroll & Swain, 1993; Ellis et al., 2006; Lado, 2008; Nagata, 1993, 1995; Nagata & Swisher, 1995; Rosa & Leow, 2004b), others have produced counterevidence or inconclusive results (e.g., Camblor-Portilla, 2006; Hsieh, 2007; Moreno, 2007; Sanz & Morgan-Short, 2004). The same can be said regarding studies which have examined the effects of either providing or not providing explicit information prior to the performance of L2 tasks: Some have found metalinguistic information to be beneficial (e.g., Alanen, 1995; DeKeyser, 1995; N. Ellis, 1993; Izumi & Lakshmanan, 1998; Robinson, 1996, 1997a; Rosa & Leow, 2004b) whereas others have not (e.g., Benati, 2004; Farley, 2004; Rosa & O’Neill, 1999; Sanz & Morgan-Short, 2004; VanPatten & Oikkenon, 1996; Wong, 2004), and potential mediating factors, such as the nature of the practice conditions and the complexity of the linguistic targets, are still being explored.

In keeping with the focus of the experiment in this dissertation, the main focus of the rest of this chapter will be on studies of computer-assisted language learning (CALL) comparing feedback that provides metalinguistic information (‘explicit’ feedback) against feedback that simply informs learners as to the correctness of their responses (often referred to in CALL as ‘implicit’ feedback). Sanz (2004) has pointed out that CALL programs provide an ideal context for investigating the effects of feedback in an individualized and consistent way. The uniformity can improve a study’s internal validity, and it can also increase the feasibility and similarity of replication studies (Rosa & Leow, 2004b).

It must be pointed out, however, that the definitions of the terms ‘explicit’ and ‘implicit’ feedback, as used in discrete-item CALL studies, differ in critical ways from how they are understood in more communicative contexts. In L2-classroom and interaction research, the term ‘explicit’ often means that attempts are made to draw learners’ attention to nontargetlike
aspects of their production in such a way that they are likely to realize they are being corrected. The term ‘implicit,’ on the other hand, means that learners are not overtly informed of the unacceptability of their utterances and therefore may not necessarily recognize any corrective intent. In contrast, in CALL research involving item-by-item practice, learners’ recognition of the corrective nature of the computer-provided feedback is often essentially taken for granted. Thus, a so-called ‘implicit’ error message in CALL (simply informing the learner of correctness) might be counted as ‘explicit’ in the context of interaction research since its corrective intent would be considered recognizable, whereas an ‘implicit’ recast in interaction might be counted as ‘explicit’ in the context of discrete-item CALL practice since it would be considered to provide additional information beyond simple acceptance or rejection of a learner’s L2 interpretation or production.

The point is that any conclusions which may be drawn about ‘explicit’ versus ‘implicit’ feedback in this review are not necessary applicable outside the realm of CALL. Excellent overviews of the extensive literature on interactional feedback exist elsewhere (e.g., Ellis & Sheen, 2006; Long, 2007; Mackey, 2007; Nicholas, Lightbown, & Spada, 2001). For studies on the effectiveness of computer-generated feedback in terms of promoting immediate uptake (as opposed to pre- to post-test improvement), see, for example, Heift (2001, 2002, 2004, 2010). For reviews of the research on feedback in the context of synchronous computer-mediated communication, see Petersen (2010) and Sauro (2007, 2009). The sections below will review studies involving discrete-item practice in which learning advantages for explicit feedback have been found, followed by studies in which advantages for explicit feedback have not been found.
Advantages found for explicit feedback

Before narrowing our focus exclusively to CALL research, it is important to mention two tightly controlled and seminal laboratory studies from the early 1990s which did not make use of computers, but in which the operationalizations of feedback types were similar to those commonly found more recently in CALL: Carroll, Swain, and Roberge (1992) and Carroll and Swain (1993). Both of these experiments determined that explicit feedback was effective, but only the latter allowed the researchers to conclude that feedback facilitated the forming of abstract linguistic generalizations.

In Carroll, Swain, and Roberge’s (1992) investigation of feedback on morphological generalizations, 79 adult English-speaking learners of L2 French at the intermediate and advanced levels were assigned either to an experimental group, in which explicit feedback was provided, or to a comparison group, in which no feedback was given. The specific target was the selectional properties of two noun-forming suffixes in French: -age, which combines productively and regularly with transitive verbs (not with true intransitives, unaccusatives, or psych verbs), and –ment, which combines productively with intransitive verbs denoting cries and sounds, and with grammatically transitive ‘pronominal’ verbs taking se (e.g., s’évanouir, ‘to disappear’ or ‘to faint’). The participants had to learn which ending to use when creating nominalizations of each kind of verb. They met with the researcher individually to complete training trials which involved being presented with sentences containing transitive or intransitive verbs and then either hearing or guessing the appropriate nominalizations. For the –age nouns, each verb was followed by a direct object in the sentence, whereas for the –ment nouns, there was sometimes an adjunct PP, but no complement. After an initial training session
with 18 items (9 with the suffix –*age*, followed by 9 with the suffix –*ment*, for each of which there were 4 models and 5 guesses, alternating in sets of 2 or 3), all participants practiced on 90 items. The experimental group received feedback on 45 items, alternating between feedback trials and guessing trials in sets of 15, whereas the comparison group never received feedback. Finally, all participants were re-tested on the 45 feedback items in a “recall” session (Recall 1), which then took place again a week later (Recall 2).

Carroll et al. found that learners in the experimental group who had received feedback tended to perform better on recall than learners in the comparison group did—“correction clearly had an effect on learning” (p. 185). However, the recall sessions involved items that the participants had already seen. Carroll et al. did not find differences between the experimental and comparison groups on their performance during the guessing sessions. Thus, they conclude that “the experimental group appeared to be learning individual items” (p. 185). Although the corrections had helped them to memorize words, there was no evidence that it had helped them to extract morphological generalizations. The authors speculate, “Perhaps generalizations cannot occur on the basis of a single exposure to 45 items” (p. 185).

In a more elaborate study focusing on the learning of dative alternation in L2 English, Carroll and Swain (1993) assigned 100 Spanish-speaking learners from low-intermediate ESL classes into four experimental groups, in which feedback of various forms was provided, and a comparison group, in which no feedback was given. Although the linguistic target involved a superficially simple binary choice of two word orders (namely, Subject + Verb + Direct Object + *to/for* + Indirect Object versus Subject + Verb + Indirect Object + Direct Object), it can be considered complex in that there are phonological and semantic constraints (e.g., involving
transfer of possession) on which verbs can participate in the alternation and which cannot (Grimshaw, 1985; Pinker, 1989). All participants met with a researcher individually, and all of the participants in the experimental groups were informed that they would be receiving feedback when they made errors. The treatment conditions differed in that some learners were given explicit metalinguistic explanations, others were simply told when their responses were wrong (i.e., given ‘implicit’ feedback, in CALL terminology), others were provided with models of the target structure when they made errors, and others were simply asked whether they were sure about their responses when they were incorrect (which may have amounted to being provided with ‘implicit’ feedback). The procedure was very similar to that followed by Carroll et al. (1992). First, there was an 8-item training session in which the participants alternated between hearing models and making guesses in sets of 2. Then, there was an experimental session with 48 items, alternating between feedback and guessing trials in sets of 12. Immediately following this, the participants were tested on 48 items in a first recall session (Recall 1), which was repeated a week later (Recall 2).

The findings, in a nutshell, were that all of the experimental groups receiving feedback outperformed the comparison group, and the group receiving metalinguistic explanations often outperformed the other experimental groups. More specifically, during the initial feedback session (i.e., during training), the metalinguistic-explanation group was already outperforming all of the other groups, and the participants who were asked about their certainty were outperforming the comparison group. During the initial guessing session (also during training), all of the experimental groups outperformed the comparison group, and the metalinguistic-explanation group outperformed the participants who were simply told when they were wrong.
or asked about their certainty. Following the training, considering just the items on which feedback had been provided, all of the experimental groups outperformed the comparison group, and the metalinguistic-explanation group outperformed all of the other groups except for the modeling group on Recall 1. Considering just the items on which feedback had not been provided, all of the experimental groups outperformed the comparison group, and the metalinguistic-explanation group outperformed the other experimental groups as well. In other words, a durable effect for feedback was demonstrated on both old (feedback) and new (guessing) items. In contrast to the conclusions of Carroll et al. (1992), the researchers were able to state that L2 learners can come to form abstract generalizations with the help of feedback, and that metalinguistic information can be particularly helpful; “simply telling a subject that he or she was wrong, providing indirect feedback, and even providing the right forms did not help as much as the explicit metalinguistic information” (p. 372).

At around the same time that Carroll and her colleagues were conducting controlled research on different types of oral feedback, Nagata (1993, 1995; Nagata & Swisher, 1995) was testing Nihongo-CALI (‘Japanese Computer-Assisted Language Instruction’), a system she had created using natural language processing (NLP) technology to provide individualized, parser-based feedback on learners’ L2 Japanese production. With this program, Nagata was able to compare two types of computer-generated feedback: (1) ‘traditional’ (T-CALI) feedback, which used simple pattern-matching techniques to inform learners about missing, inaccurate, and unexpected words and particles in their sentences (Nagata, 1993; Nagata & Swisher, 1995), along with information regarding the errors’ locations (Nagata, 1995); versus (2) ‘intelligent’ (I-CALI) feedback, which, in addition to doing the above, used more sophisticated NLP techniques
to provide learners with detailed metalinguistic information about the nature of their errors.

Both feedback types were explicit, but the I-CALI type was more informative.

The participants in Nagata’s first study (reported in Nagata, 1993, and Nagata & Swisher, 1995) were 32 university students taking a second-year course in Japanese as a foreign language. They were assigned in matched pairs to the T-CALI and I-CALI conditions on the basis of a written grammar test. The linguistic targets were three types of passive structures (direct, adversative, and honorific) and postpositional particles, which are used to indicate the grammatical and semantic roles of noun phrases in Japanese. A 20-minute explanation of the passive structures, provided to both groups in class, was followed by four CALI sessions, the first three of which each targeted a different passive structure, and the fourth of which reviewed all three. All participants were given the same grammar notes to read for about 7 minutes at the beginning of each hour-long session, and they could refer back to the notes whenever desired. The exercises (90 items in total) involved reading sentences produced by imaginary Japanese conversation partners and then responding to them in writing. Following the four sessions, the participants took a 20-item achievement test in class and completed a 28-item questionnaire with Likert-style items regarding their attitudes toward the computer activities. Three weeks later, after an additional 30-minute recitation session in which passive structures were practiced, the participants took an exam on which 4 of the questions involved passives. The only manipulated difference between the T-CALI and I-CALI groups was in the content of the feedback they received.

Comparing the groups’ post-test performance, Nagata and Swisher (1995) found that the I-CALI learners, who had received more informative feedback, scored statistically higher
than the T-CALI learners. Whereas the two groups appeared to make roughly the same numbers of mistakes with lexis and verb morphology, the T-CALI group seemed to make more errors with particles. The researchers explain this finding, convincingly, by noting that the learners had been advised as to which passive construction they should use for each exercise. Since forming a passive in Japanese requires simply attaching a suffix to a verbal stem and the learners had already largely mastered conjugation issues related to verb tense/aspect, the less informative T-CALI was likely sufficient in that area. Regarding vocabulary, considering that the learners had been given access to an online dictionary, both types of feedback essentially provided everything they needed to know simply by indicating which lexical items were wrong or missing. Particles, on the other hand, are exceedingly complex and can continue to pose problems for learners of Japanese even at advanced stages.

Analyses of the questionnaire data suggested that the I-CALI learners found the feedback somewhat overwhelming at times, but easier to understand and more useful in helping them to comprehend why their answers had been wrong. In contrast, the T-CALI feedback sometimes confused the learners and did not adequately facilitate correction of their particle errors. According to Nagata and Swisher, merely drawing attention to the fact that certain particles were wrong, missing, or unexpected was not sufficient to help learners understand exactly how or why their answers were nontargetlike.

In a very similarly designed study, Nagata (1995) continued with this line of research, improving on the less-informative T-CALI feedback by augmenting it with information regarding the locations of learners’ errors. The participants were 18 university students in a first-semester Japanese course, and the linguistic target involved postpositional particles, which in this case
were quite new to the participants. Assigned to T-CALI and I-CALI groups in matched pairs on the basis of midterm scores, the participants engaged in six experimental sessions over the course of two weeks. First, they were given a 30-minute in-class introduction to particles along with a pre-test (10 fill-in-the-blank and 2 sentence-construction items). This was followed by four CALI sessions, the first three of which focused on three particle functions each, and the fourth of which reviewed all nine functions, with a total of 67 items. As before, all participants were given grammar notes to read at the beginning of each session, and each exercise involved reading a communicative context and responding to an imaginary Japanese conversation partner. Following the CALI sessions, the learners completed an in-class post-test which included the same questions they had been given on the pre-test, plus an additional 10 fill-in-the-blank and 9 sentence-construction items. They also filled out an attitudes questionnaire. Again, the only manipulated difference between the T-CALI and I-CALI groups was in the content of the feedback they received.

Nagata (1995) found that the more detailed I-CALI feedback was more effective; although both groups showed statistically significant improvements from pre- to post-test, the I-CALI group demonstrated greater gains. Both groups had positive attitudes toward the computer program. On several of the questionnaire items, the average ratings (on a scale from 1 to 5) were very similar (e.g., 4.8 for both groups in response to the statement I learned a lot from the CALI exercises; 4.6-4.7 for I find the CALI exercises interesting). The items where T-CALI and I-CALI participants’ ratings tended to differ a bit more (though not statistically significantly) were for statements such as The error messages are helpful in pointing out why my response is
wrong (I-CALI 4.6 vs. T-CALI 3.8) and I want to practice Japanese by using this type of CALI exercise on a regular basis (I-CALI 4.9 vs. T-CALI 4.1).

In a study that was more complex in terms of research design, Rosa and Leow (2004b) conducted a pre-test/post-test/delayed-post-test experiment comparing the effectiveness of treatment conditions with varying types and degrees of explicitness. The participants were 100 English speakers learning L2 Spanish in fifth-semester university courses, and the linguistic target was past counterfactual conditionals, a structure that can be considered morphologically and semantically complex due to the multiple agreement, tense, and mood morphemes used to express abstract re-imaginations of events. Six groups were created by combining the variables [+/- task essentialness], [+/- explicit pre-task], and [explicit vs. implicit feedback]; randomly assigned thus to different treatment conditions, the participants were provided either with (1) an explicit pre-task and explicit feedback, (2) an explicit pre-task and implicit feedback, (3) an explicit pre-task only, (4) explicit feedback only, (5) implicit feedback only, or (6) the same sentences to read for meaning without engaging in the computerized experimental task.

The explicit pre-task, which took 8 minutes, included an introduction to the concept of counterfactuality, an overview of the verb tenses used in present and past counterfactuals, and glossed examples with interactive multiple-choice questions and explanations. The implicit feedback simply informed the learners about the correctness of each answer, whereas the explicit feedback included an explanation of why each answer was right or wrong. The task itself involved selecting from four possible choices an appropriate subordinate clause (i.e., with the appropriate verb tense/mood) to complete a past counterfactual conditional whose consequent was provided. Each item was presented in the form of a computerized multiple-
choice jigsaw puzzle with moveable elements. After making each selection, the participants wrote down their responses. All produced concurrent think-aloud protocols while performing the 28-item task, for which they were allowed 28 minutes. Ten of the items involved present counterfactuals, while 18 were past counterfactuals (the target). The learners’ recognition and production of the target were tested with both old and new exemplars. The recognition tests were given in a multiple-choice format that involved deciding which of 4 verb forms best completed the subordinate clause of a conditional statement which was provided. The production tests were given in a fill-in-the-blank format that similarly involved conjugating a verb provided in parentheses in order to complete a conditional’s subordinate clause.

Rosa and Leow (2004b) found that all of the experimental groups showed statistically significant improvement on the recognition and production tests and that, except for the implicit feedback group’s production of new exemplars, all of the experimental groups outperformed the control group on all measures. On the recognition of new items and production of old items, the implicit feedback group was outperformed by the learners who had received explicit feedback alone and by the learners who had been given an explicit pre-task plus feedback of either kind. On the production of new items, the implicit feedback group was again outperformed by the learners who had been given two sources of information. There were also trends suggesting that the explicit pre-task group might have been outperformed by the learners who had been given two sources of information, and this was a statistically significant difference for the most explicit group on the production of new items. Also potentially of interest is that the only experimental group that consistently maintained its gains on the delayed post-test was the one in which participants had both completed the explicit pre-
task and received implicit feedback. The participants who had been given a single source of information (i.e., either the explicit pre-task or explicit feedback or implicit feedback) showed a decrease in the accuracy of their production of old and new items, and the explicit feedback group also showed a decrease in the accuracy of its recognition of old items. The learners who had both completed the explicit pre-task and received explicit feedback showed a decrease in production of old and new items as well as a decrease in recognition scores on new items.

The authors conclude that although exposure to L2 input with implicit feedback did lead to improvement, “higher degrees of explicitness had a more drastic impact” (p. 210), and although two sources of information appeared to produce the best results, “when only one source was available, the most advantageous method seemed to be explicit feedback” (p. 209). The explicit pre-task did not seem to be an essential component considering that learners given only explicit feedback “derived an amount of generalizable knowledge comparable to that of learners receiving feedback and a pretask” (p. 209), whereas learners given only the explicit pre-task never outperformed those receiving only implicit feedback.

Among the limitations of this study, Moreno (2007) has noted that the testing and treatment tasks were highly controlled in nature, requiring only the selection or production of an appropriate verb form. Moreover, since the present counterfactual (distractor) test items were not included in the scoring, the researchers may have been ignoring the important issue of whether the participants knew how to choose between present and past counterfactuals. Sanz and Morgan-Short (2004) have additionally pointed out that the limited number of training items may have disadvantaged the groups receiving less explicit information since implicit learning is slow and can require a great deal of exposure (N. Ellis, 1993, 2005).
Therefore, in drawing conclusions based on the results of the study, it may be prudent to replace phrases such as “more drastic” and “advantageous” with the phrase “more efficient”.

Similar findings, as well as similar caveats, were the outcome of a computer-based study by Bowles (2005), whose results also suggested that more explicit feedback might be more efficient. In this experiment, 150 English-speaking learners of L2 Spanish in first-semester university courses were randomly assigned into six treatment conditions which varied in terms of type of feedback (explicit versus implicit) and verbalization (non-metalinguistic versus metalinguistic versus silent). The linguistic target was dative experiencer constructions with the verb *gustar* (‘to be pleasing to’), frequently a source of mistakes for native speakers of English, who have a tendency to confuse semantic roles in relation to its non-canonical (OVS) word order (e.g., *Me₁ gusta₂ el campo₃* = ‘The countryside₃ is pleasing₂ to me₁’). As in Rosa and Leow’s (2004b) experiment, the treatment sessions were short, lasting less than an hour and containing only 24 exemplars of the target structure, potentially disadvantaging the learners who received less explicit information. Also as with Rosa and Leow’s study, researchers have raised concerns regarding the controlled nature of the production task, which involved translation from English, and the scoring procedure employed, in which learners were not penalized for certain crucial types of mistakes. Lado (2008) has pointed out, for example, that impossible SOVS sentences such as *Yo me gusta el campo* (literally, ‘I to me the countryside is pleasing’) were marked correct despite the fact that the inclusion of a doubled subject in preverbal position revealed that the participants had not internalized “the key structural reason behind the form” and might have been conceptualizing it incorrectly as SVO: *Yo (‘I’) me-gusta (‘like’) el campo (‘the countryside’)* (p. 59).
Before engaging in the experimental tasks, all participants were given a computerized tutorial with practice exercises regarding Spanish indirect object pronouns (e.g., me, ‘to me’). They then completed a series of mazes in which, step by step in order to proceed, they had to choose the appropriate Spanish words to create the equivalents of English sentence that were provided for them. In the implicit feedback condition, choosing an incorrect sentence element resulted in an immediate dead end. In the explicit condition, in addition to this sort of task-inherent feedback, the learners were given fairly extensive metalinguistic information at each choice point, regardless of whether the choice had been correct or incorrect, explaining what they might have thought based on the way English works and also explaining the right way of thinking about the various sentence elements in Spanish. In other words, independently of whether the participants in the more explicit group apparently needed feedback, they were constantly provided with detailed explanations of the linguistic target.

Examining the participants’ immediate post-test performance (also conducted as a computerized maze), Bowles (2005) found that, for both old and new items, learners who had received explicit explanations outperformed those who had received implicit feedback. However, by the time of the delayed post-test, which was administered three weeks later in a paper-and-pencil format with the same items in a different order, the advantage for explicit feedback had disappeared; the mean score for the explicit group was statistically similar to that of the implicit group. Thus, although explicit explanations may have been more efficient in terms of producing immediate gains following limited exposure, not all of the explicit group’s knowledge appeared to have been retained over time when tested in a different format, and
concerns about the study’s scoring procedures mean that the findings should be interpreted with caution.

A final computer-based study which found qualified advantages for explicit feedback was conducted by Lado (2008). In addition to investigating the relative effectiveness of two types of feedback in the *ab initio* learning of Latin as an L3, Lado explored relationships with individual differences in the learners’ levels of English-Spanish bilingualism, working memory capacities, and language aptitude. The linguistic target was the use of case morphology to assign semantic roles to noun phrases in Latin sentences, and the participants were 151 basic, intermediate, advanced, and native-like speakers of L2 Spanish with English as their L1. At the first session, they completed a Latin vocabulary-learning activity and vocabulary quiz, followed by 4 pre-tests (written and aural interpretation of Latin sentences, grammaticality judgments, and production of new sentences based on pictures) and a working memory test. At a second meeting, which was held about a week later, the participants engaged in 2 practice sessions, each of which involved 6 different tasks (3 written, 3 aural) with 9 or 10 items each (112 items total). Then, they completed 4 post-tests in the same format as the pre-tests and took tests of their phonological short-term memory capacity. Two weeks later, at a third meeting, they completed 4 delayed post-tests and the Modern Language Aptitude Test.

The computerized treatments involved input-based practice with feedback which informed the participants about the correctness or incorrectness of their responses. In one of the two experimental groups, explicit feedback was additionally provided in the form of metalinguistic explanations. During these treatments, all participants practiced choosing from 2
options the correct English translation or the appropriate picture to match a written or aurally presented Latin sentence, and choosing which of 2 Latin sentences best represented a picture.

Lado (2008) found that the learners who had received explicit feedback demonstrated more accurate immediate post-test performance than those who had received only feedback on the correctness of their answers. However, this advantage did not last to the delayed post-tests on all measures; at that point, although an advantage for the explicit condition was still found on written interpretation and production, the groups performed similarly on the grammaticality judgments and aural interpretation of new items. Lado suggests that possibly “implicit treatments may require more exposure to the items or to the feedback in order for output advantages to appear” (p. 243). A possible limitation of the study that Lado mentions is that since the feedback appeared on the computer screen for only 5 seconds, the participants in the less explicit group may have needed more time for conscious reflection and hypothesis testing, or more instances of exposure, if they were trying to generate rules.

**No advantages found for explicit feedback**

In several of the studies discussed above, explicit feedback was found to be more beneficial than implicit feedback, but the advantages were not always maintained over time. Several other studies have found no advantages for explicit feedback. In a computer-based experiment by Sanz and Morgan-Short (2004), for instance, participants who had received no metalinguistic information were found to perform indistinguishably from participants in other experimental groups who had been given explicit prior explanations, explicit reactive feedback, or both. Their study examined 69 L1-English/L2-Spanish learners’ ability to use case morphology
to assign semantic roles correctly in non-SVO sentences with O-clitic V structures (e.g., *Lo besa la chica*, ‘Him kisses the girl’ = ‘The girl kisses him’). The experimental groups were created by combining the variables [+/- prior explanation] and [+/- explicit feedback].

The treatment activities, which included 56 items in total, involved aural and written interpretation at the sentence level (identifying which of 2 pictures represented the meaning of a given sentence) and written interpretation at the text level (e.g., reading a short article and deciding which of 2 titles was appropriate). All participants were given at least implicit feedback on each response (i.e., either “OK” or “Sorry, try again”). In the explicit feedback condition, following incorrect answers, the participants were additionally provided with explanations that included metalinguistic information about the target structure, how to process it in Spanish, and the consequences of applying the wrong processing strategy.

The tests involved both interpretation and production. On the 15 interpretation items (10 of which involved the linguistic target), participants had to choose which of 2 pictures represented the meaning of a sentence. One of the production tests was similar: For each of 15 items (11 targets), the participants had to fill in a blank with the appropriate pronoun to express who was doing what to whom in a given picture. The other production test involved watching a silent video portraying several interrelated events and then retelling the story in writing, which naturally called for the use of pronouns.

Sanz and Morgan-Short (2004) found that all of the groups showed statistically significant improvement on both the interpretation and the production tests, with no statistically significant differences among the conditions. In view of this, they conclude that “explicit information plays a minor role at most” in the context of structured-input activities in
which participants are required to process the critical forms for meaning (p. 45): “With certain
tasks,” they suggest, “the task itself and not the explicit condition under which the task is
carried out may lead to acquisition” (p. 43). The authors do acknowledge, however, that they
did not compare the implicit feedback condition against a condition in which learners received
no information regarding the correctness of their responses. As such, they admit that they
cannot say for certain that the structured input alone produced the benefits.

Sanz and Morgan-Short explain that the differences between their results and those of
other studies could have been due to the number of items, the number of choices, or the
degree of task-essentialness in the implicit feedback condition (e.g., Rosa & Leow, 2004b), or to
the complexity of the linguistic targets (e.g., Nagata, 1993, 1995). The testing and treatment
items always involved binary choices, and feedback on the correctness of the participants’
responses, at a minimum, was always provided during the treatments. In relation to this
constant provision of binary feedback, it is worth considering that the information that a
particular form does not have a particular interpretation could, in itself, potentially promote a
metalinguistic stance. Moreover, if a linguistic target is straightforwardly understandable, as
seems to have been the case here with the distinction between subject and object case
morphology, then the equivalent of explicit metalinguistic information may be so easily
inferable (at least for some learners) as to be virtually present (Rosa & O’Neill, 1999).

The authors acknowledge this, stating that “when there are a limited number of
possibilities in the input and the number of responses is also limited, locating the source of the
problem and rejecting hypotheses is possible even without explicit feedback” (p. 73). They
additionally note that the linguistic target in their study may have been quite salient to the
participants given that it was presented in sentence-initial position, in sentential and suprasentential contexts. Considering that feedback was provided only as needed, Moreno (2007) would add, it may be that participants in the explicit feedback group were not exposed to much metalinguistic information in Sanz and Morgan-Short’s study.

An experiment by Moreno (2007), in fact, corroborates Sanz and Morgan-Short’s (2004) findings with the same linguistic target (preverbal direct object clitics), but employing a task more similar to that used by Bowles (2005). In Moreno’s computer-based experiment, 57 English-speaking learners of L2 Spanish in beginning-level university courses were randomly assigned to four treatment conditions that differed with respect to the variables [+/- task essentialness] and [+/- explicit feedback]. All of the participants produced concurrent think-aloud protocols. The research question relevant to the present discussion was whether the type of feedback would have a differential effect on the learners’ recognition and oral and written production of the targeted structure. The treatment task involved describing 16 pictures in which one of the people depicted was performing an action that clearly affected another person. As in Bowles’ (2005) experiment, learners in the task-essential conditions played a game that involved making linguistic choices to travel down a path. When they chose the correct object clitic to describe the picture, the trip continued and the computer completed the sentence for them, adding the verb and subject. In the implicit feedback group, learners were informed only about the correctness of their choices. In the explicit feedback group, they were given a metalinguistic explanation every time they clicked on an answer in order to ensure that all of them received the same amount of information. In both groups, when the learners’ answers were wrong on a first attempt, they were prompted to try again. After that, the
computer provided the correct answer. The tests involved choosing which sentence could be used to describe a picture (given 3 different options or ‘none of the above’), or producing, orally or in writing, a sentence to describe a picture.

The very low number of treatment items in Moreno’s (2007) study might be considered a serious limitation but for the fact that it allowed her to call previous researchers’ arguments into question and shed light on the results of earlier studies. Not only did Moreno find no evidence that explicit feedback was more beneficial than implicit feedback over the short duration of her experiment, but she found some evidence suggesting that implicit feedback may have been more effective. Both groups demonstrated statistically significant gains from the pre-test to the delayed post-test, and there were no statistically significant differences between them on the immediate post-test. However, except in the implicit group’s oral production performance, participants showed a statistically significant decrease in scores from the post-test to the delayed post-test, and the learners who had received implicit feedback actually outperformed the explicit group on the delayed post-test of oral production. In a sense, this echoes the results of other studies in which explicit feedback groups have not always completely maintained their gains or differences from implicit groups over time (e.g., Bowles, 2005; Lado, 2008; Rosa & Leow, 2004b).

Weighing her results along with those from previous CALL studies, Moreno points out that there seems to be a trend for “implicit feedback [to be] just as efficient as explicit feedback when students of lower proficiency levels learn with simpler structures, while explicit feedback appears to be more effective with more complex structures and higher proficiency levels” (p. 164). In relation to Sanz and Morgan-Short’s suggestions as to why their results may have
differed from Rosa and Leow’s (2004b) findings, Moreno notes that her study’s treatment session provided participants with even fewer instances of exposure to the linguistic target (16) than Rosa and Leow’s (18); nonetheless, there were still indications of an advantage for the implicit group. She argues that perhaps, “at least at the beginner’s level, a massive amount of exposure is not necessary if each occurrence is designed in such a way that not too many of the learner’s attentional resources are exhausted in finding the correct answer” (pp. 153-4).

Regarding the issue of linguistic complexity, Lado (2008) provides an insightful perspective on how superficial similarities may mask important differences. In her experiment as well as in those performed by Sanz and Morgan-Short (2004) and Moreno (2007), the participants had to learn to assign more weight to a morphological cue and less weight to a syntactic one; however, whereas the learners in Sanz and Morgan-Short’s study simply had to become familiar with the flexibility of word order in Spanish and then attend to 4 different morphological forms, in Lado’s study, the learners had to adjust to the fact that case assignment in Latin was more reliable than either word order (as in their L1, English) or subject-verb agreement (as in their L2, Spanish) and then attend to 8 different morphological cases, all of which were new to them as beginning learners of Latin. “Hence,” she concludes, “from the results obtained in the majority of CALL studies… it appears that providing metalinguistic rules gives an immediate advantage when the target form is complex” (Lado, 2008, p. 239).

A final CALL study in which the simplicity of the target form may have contributed, at least in part, to a lack of differences between feedback conditions was Camblor-Portilla’s (2006) experiment investigating the effectiveness of implicit, explicit, and interactive feedback on noun-adjective agreement in L2 Spanish. The gender and number morphemes involved in this
linguistic phenomenon tend to be a source of mistakes for L1 English speakers; still, compared to the difficulties involved in choosing postpositional particles (Nagata, 1993, 1995), forming past counterfactual conditionals (Rosa & Leow, 2004b), and so on, it is fairly straightforward to mark agreement on adjectives which are adjacent to the nouns they describe.

In Camblor-Portilla’s (2006) study, 77 English-speaking learners of L2 Spanish in first-year university courses were randomly assigned to three feedback groups and a comparison group. After they first completed a lesson on vocabulary, the roughly 30-minute treatment task involved writing descriptions of 24 pictures of objects of various colors. Spanish glosses were provided, showing the appropriate gender of each noun. In the implicit feedback condition, any mistakes the participants made in noun-adjective agreement were capitalized, whereas in the explicit feedback condition, metalinguistic information was provided in addition to this. For participants in the interactive feedback condition, besides having their mistakes capitalized, they played a matching game in which, to describe a picture, they had to choose a noun in either the singular or plural form and then decide on the appropriate form for the adjective (feminine singular, feminine plural, masculine singular, or masculine plural). Immediately after having received feedback on a sentence, learners in all three conditions had to rewrite it. The comparison-group learners also had to rewrite their sentences making any changes they deemed necessary, but they were never given feedback.

Camblor-Portilla found that the learners who had received feedback performed better than those who had not, but she did not find any conclusive results regarding the relative benefits of different types of feedback. In fact, noting that learners in the comparison group also showed statistically significant improvement over time, Camblor-Portilla suggests that,
with such an easy and familiar target, merely completing the task itself may have been somewhat helpful.

**Learner awareness of information provided in feedback**

Many studies of L2 feedback, particularly those comparing more and less explicit types, invoke Schmidt’s (1990, 1995, 2001) noticing hypothesis, which states that “SLA is largely driven by what learners pay attention to and notice in the target language input and what they understand the significance of noticed input to be” (2001, p. 4). Schmidt defines ‘noticing’ as the conscious registration of “elements of the surface structure of utterances in the input—instances of language, rather than any abstract rules or principles of which such instances may be exemplars” (p. 5). He accepts Robinson’s (1995) construal of noticing as being equivalent to the detection of language forms plus the rehearsal of them in short-term memory. Importantly, Schmidt entertains the hypothesis that “nothing is free…. learners must attend to and notice any source of variation that matters, whatever makes a difference in meaning” (pp. 6-7), and the “requirement of noticing is meant to apply equally to all aspects of language” (1990, p. 149):

In order to acquire phonology, one must attend to the sounds of target language input, especially those that are contrastive...if one’s goal is to sound like a native speaker, one must attend to sub-phonemic details as well.... In order to acquire pragmatics, one must attend to both the linguistic form of utterances and the relevant social and contextual features.... In order to acquire morphology (both derivational and inflectional) one must attend to both the forms of morphemes and their meanings.... [etc.] (2001, pp. 30-31).
In non-feedback research, through the use of concurrent think-aloud protocols, a great deal has been discovered about learners’ attentional processes during L2 task completion. Several studies have found that reports of awareness of L2 forms at the level of noticing are associated with L2 development, and reports of awareness at the higher level of understanding (e.g., containing rules about the linguistic targets) are even more strongly associated with demonstrations of learning (e.g., Leow, 1997, 2000, 2001; Rosa & O’Neill, 1999). Researchers have noted that meta-awareness on the part of learners often involves rather sophisticated conceptually driven processing, such as hypothesis-testing approaches to L2 data that “allow for extraction of patterns from specific instances” (Rosa & Leow, 2004a, p. 287).

In the computerized feedback study by Rosa and Leow (2004b) already discussed, corroborating results by Rosa and O’Neill (1999), the researchers found that “learners processed input with various degrees of awareness of the target structure as a function of the explicitness of each learning condition” (Rosa & Leow, 2004a, p. 287). The learners who had received feedback during task completion were more likely to engage in hypothesis formation than those who had been given only the explicit pre-task, and none of the learners in the comparison group (who failed to show statistically significant improvement over time) formulated hypotheses about the target structure. Having a greater number of sources of information about the linguistic target was associated with higher frequencies of reports of awareness at the level of understanding, and when only one source of information was available, the explicit sources of information (i.e., the explicit pre-task instruction and the explicit feedback) were more likely to be associated with higher levels of awareness than the implicit feedback was.
Although similar in many respects, Camblor-Portilla’s (2006) study produced somewhat different results. In that experiment, learners in the three feedback conditions (implicit, explicit, and interactive) were more likely to report awareness about the linguistic target than those in the comparison group were. However, clear differences did not emerge among the three feedback conditions; “all experimental groups reported a relatively similar level of awareness” (p. 194), possibly “as a consequence of the fact that gender, and especially number agreement, was not a completely unknown concept to most of the learners” (p. 191). Interestingly, a qualitative analysis of the learners’ verbalization data suggested that, as opposed to leading to differences in processing, the different types of feedback simply “extended or reduced the process of achieving a relatively high level of awareness” (p. 207). Even learners in the implicit feedback condition eventually tended to discover a rule for noun-adjective agreement.

Rosa and Leow (2004a) point out that researchers have traditionally only been able to infer that feedback encourages learners to attempt to extract generalizations; thus, their work makes an important contribution since it provides “empirical evidence of the types of cognitive processes activated through feedback during on-line input processing” (p. 288). Their study also sheds a bit more light on the issue of maintenance of gains over time in relation to learner awareness. Participants who had reported awareness at the level of noticing and those who had reported awareness at the level of understanding showed statistically significant improvement over the course of the tests; however, those in the latter group, with a higher level of awareness, also showed a statistically significant decrease in scores from the immediate to the delayed post-test, a result which resonates with findings from several other studies to the effect that learners in more explicit treatment conditions, although usually maintaining
their performance at the level of those in less explicit conditions, may not retain all of their improvement over time (e.g., Bowles, 2005; Lado, 2008; Moreno, 2007).

It may be, as Schmidt (2001) proposes, that there is an advantage for higher levels of awareness in the *initial recognition and analysis of L2 features*. Input and intake are products of language processing; what is salient and what amounts to ‘noise’ for a learner depends in part on existing mental representations and parsing procedures (see, e.g., N. Jiang, 2004, on morphological insensitivity; VanPatten, 2007, on processing preference principles; MacWhinney, 2011, on cue strength; etc.). Thus, the simple noticing of relevant features may not always be likely: “What is noticed... [is] input as interpreted by existing schemata” (2001, p. 31). Schmidt speculates that there are a variety of circumstances in which awareness at the higher level of understanding may sometimes be necessary—for example, “when the target language requires that sources of information be attended that are not attended in the L1” or “when information that is automatically processed in the L1 (without reaching awareness) must be suppressed or treated differently in the L2” (p. 29). More focal attention may also be required, according to Schmidt, when long-distance linguistic relationships are involved, and, perhaps, “getting things right when key elements are not overtly expressed will be most difficult” (pp. 24-5).

Considerations such as these make it important to explore under what circumstances learners can detect errors and assign blame, which classes of linguistic forms may be more or less amenable to awareness at the levels of noticing and understanding, and how different mental systems (e.g., underlying explicit and implicit knowledge and learning; involving different aspects of language, such as semantics and morphology, etc.) can and cannot work together. Along these lines, there is an interesting proposal, by Carroll (2001), that has not yet
been explored empirically in research on L2 feedback, whether communicative or lab-based: namely, that certain areas of language, due to the architecture of the language faculty, might be *a priori* impermeable to the influence of feedback or metalinguistic information.

**Carroll’s ‘awareness constraint on negative evidence’**

S. E. Carroll (2001) argues persuasively that, in order to be truly explanatory, a theory of SLA must include several kinds of theories. She agrees with Gregg (1993, who cites Cummins, 1983) in his contention that explanations of language acquisition require both a property theory of linguistic knowledge (i.e., learners’ mental representations) and a transition theory of how such knowledge can change (i.e., how mental representations can restructure in principle). However, for Carroll, “property and transition theories are still only one-half of the story” (p. 38). In addition, she would contend that the field of SLA also requires a performance theory of linguistic processing (i.e., how learning mechanisms gain access to stimuli or conceptual information to reanalyze and encode grammatically), and a learning theory of how new linguistic knowledge is created (i.e., how restructuring takes place given the availability of new information to the learning mechanisms).

In Carroll’s own Autonomous Induction Theory (AIT), the language faculty is claimed to have a specific structure, realized in autonomous representational/processing systems. As Carroll (2007) explains it, this simply means that, for example,

> [D]istinctions encoded in a phonetic representation (adverting to constructs like vocal fold vibration, bursts of noise, wave duration, or wave amplitude) are quite different from the distinctions encoded in a conceptual structure (adverting to
The different types and levels of representation have unique sorts of primitives and principles of combination which are not always in one-to-one correspondence with each other. Therefore, in order for them to be able to interact, there must be ‘correspondence rules’ (Jackendoff, 2002) which can translate information between different formats. Since “not all types of information represented in one representational format can be ‘redescribed’ in other representational formats” (Carroll, 2001, p. 367), information can be lost when the modular systems interact—a fact which, she argues, has important implications for learning.

Regarding explicit forms of learning mediated by conscious thought processes, Carroll claims (and considers it especially relevant) that there are “severe limitations” on how information encoded in conceptual structures can interact with the grammatical representations involved in parsing (p. 120). For certain linguistic phenomena, there simply are no concepts that map onto them precisely. In the case of grammatical gender, for example, “no relevant category can be induced from conceptual categories” (p. 199). Furthermore, based on Jackendoff’s (1987, 2007) Intermediate Level Theory of Consciousness, Carroll assumes that awareness is modality-specific and, crucially, that “the only level of representation which presents itself to awareness is the phonological level” (Carroll, 2001, p. 127). Under this assumption, there are features of language (e.g., semantic, phonetic) that speakers can never be aware of, given that they “are not re-representable in phonological representations” (p. 27).

As far as feedback is concerned, Carroll maintains that “empirical evidence... shows that feedback and correction play a causal role in SLA... [however,] this role is limited to specific
kinds of language acquisition problems” (p. 2). She assumes (not uncontroversially) that
negative evidence requires a metalinguistic interpretation in order to affect L2 knowledge: A
learner must understand that s/he is being corrected, attend to a particular aspect of the
utterance, develop a more detailed representation of its structure, detect an error, and assign
blame. Combining this assumption with her assumptions about the nature of awareness, Carroll
reasons that the relevant element(s) or distinction(s) must be instantiated somehow in
phonological form in order for a learner to be aware of them and for the feedback to be
effective. As Carroll puts it,

To the extent that feedback, correction, and metalinguistic instruction depend on
awareness, they cannot have any effect on the acquisition of grammatical
information available exclusively to the morphosyntactic representations, or to
information in conceptual structures. I will call this the awareness constraint on
negative evidence.... [T]he only levels of representation that either we
[interlocutors] or they [learners] can be aware of are phonological
representations (p. 128).... Since the feedback is encoded in conceptual
structures, but the grammar is encoded in a variety of other autonomous
systems, there must be points of interaction where the information in the
conceptual structures can be re-encoded in the proper format... Where no
correspondences exist, feedback will have no effect (p. 203).

Carroll points out explicitly that her theory is different from theories claiming that negative
evidence and metalinguistic information cannot affect language acquisition (e.g., Schwartz,
1993; Truscott, 1998, 1999). Rather, she believes that “the investigation of exactly what the
effects of feedback and correction are can shed light on the nature of modularity and information processing” (p. 50).

What exactly does it mean to say that a learner can be aware only of phonological representations? Some examples given by Carroll (2001) include “lexical information – in the form of syllables, feet and prosodic words” (p. 372); “stress shifts, order of syllables, rhymes, and differences in the tones of intonation contours or tonated words” (p. 349); “the ordering of lexical items [or prosodic constituents] in phonologically realized s-structures” (p. 357); and “properties like membership in a given word class, semantic roles or semantic role and case assignment functions... to the extent that these can be directly mapped onto the phonological representations of a word...for example...that sink but not think belongs to the same conjugation class as drink” (pp. 129-30).

In contrast, learners cannot be aware of the contents of phonetic, morphosyntactic, and semantic representations. For example, “no amount of being told ‘It sounds like [...]’ should be able to influence our perception and learning of the acoustic features of acoustic-phonetic stimuli (more specifically, the details of intonation, stress, or segments) independently of what our perceptual systems will process bottom-up anyway”—i.e., “lower levels of parsing” (p. 129). Learners also cannot be aware of “case or word class [when it] cannot be instantiated in specific phonological distinctions” (p. 130). From the perspective of a learner’s interlocutor, it is not possible to be aware of “the morphosyntactic contents of the learner’s representation of the erroneous sentence (the degree of embeddedness, the categorial features, the improper binding of an antecedent-trace)” (p. 128). The construct of embedded clauses should be accessible only via linear notions such as precedence since “[the former] construct is eliminated
in the transition from s-structures to phonological structures” (p. 388). “[T]he leftward or rightward direction of syntactic attachment” also is not learnable via feedback, “although the leftward or rightward direction of cliticization might be learnable through phenomena like ellipsis, deletion, and so on which involve the omission of prosodic words” (p. 129). In sum, “aspects of meaning, morphosyntax, internal morphological structure of words, sub-segmental aspects of phonology and all of phonetic knowledge” are predicted to be impervious to feedback and correction (p. 394). Carroll’s own summary of her predictions runs as follows:

(a) Phonetic representations should be completely impermeable to feedback and correction.

(b) The phonological information stored in lexical entries, i.e. the linear order of segments and syllables, but not the content of segments, should be permeable to feedback and correction.

(c) Feedback, correction, and metalinguistic information should be able to influence how we hear integrated phonological representations.

(d) They should not be able to influence the linear and hierarchical patterns of s-structures except insofar as these are directly encoded in phonological representations.

(e) They should be able to influence the morphosyntactic information stored in lexical entries, e.g. s-selection and c-selection properties of individual words to the extent that these are associated with phonological forms.

(f) They should not be able to influence semantic properties of linguistic units except insofar as these are instantiated in phonological forms (pp. 128-129).
If some of these predictions do not intuitively seem plausible, it may be in part because, from a weak-interface perspective (e.g., N. Ellis, 2005), the assumption is not that conceptual representations must be transformed somehow into functioning grammatical elements. Rather, an idea is that conceptual representations can help learners to interpret (or produce) sentences in more targetlike ways and thereby provide better L2 input for themselves. Moreover, even if it is assumed that linguistic awareness is mediated through phonological representations, it is also true that humans experience the world consciously through a variety of other modalities which produce other sorts of phenomenality. Jackendoff would not deny this; his Intermediate Level Theory of Consciousness was partially inspired by Marr’s (1982) theory of vision, and Jackendoff (2007) points to Prinz’s (2007) compilation of evidence from neuroscience supporting the idea that consciousness arises at intermediate levels of perceptual hierarchies in other senses as well. Meanwhile, there is evidence to believe that working memory, generally recognized as being closely connected with consciousness, has a multi-componential structure which includes, among other functions, a phonological loop, a visuo-spatial sketchpad, and an episodic buffer in which information from different modalities is integrated to create new representations (Baddeley, 1986, 2003, 2007). Emphasizing the coherence of conscious experience, Baars (1988, 1997) has used metaphors of a global workspace and a theater stage to suggest that consciousness can act as a forum for different sorts of information to be activated simultaneously and for new associations to be created.

These considerations make it interesting to conjecture that there may be circumstances in adult SLA under which the presentation of information in different representational formats could facilitate the effectiveness of feedback. For instance, visual presentations of articulatory...
gestures might help learners to have conscious oral proprioception of sub-phonemic features (e.g., [+rounded]) which are important in the L2 but not in the L1 and which they might not otherwise attend to (see, e.g., research by Hardison [2003] showing that auditory + visual training is more effective than auditory training alone for adult L1-Japanese and L1-Korean speakers learning to discriminate /r/ versus /l/ in L2 English). Visual depictions of paths and manners of motion might help learners to be aware of at least certain semantic features which are relevant to understanding how particular lexical items in the L2 work. Linguists likely ‘experience’ c-command as a hierarchical relationship more readily via tree diagrams than with linear brackets. As Baars (1997) puts it, “Even the most abstract concepts may have qualitative mental ‘access images’ of some kind” which can make them easier to understand and manipulate (p. 81). To illustrate, he gives the example of visualization in geometry, which has had a large impact in mathematics despite the fact that geometrical figures can also be expressed algebraically. The point is simply that it may sometimes be possible to augment and enhance the representational formats L2 learners have at their conscious disposal. If this can lead to more targetlike interpretations of L2 data, perhaps it can facilitate SLA.

Before taking steps in this direction, however, it is important to consider whether the L2 feedback research that has already been conducted has anything to say empirically about Carroll’s predictions. Table 1.1 presents a representative sample of research investigating the effectiveness of feedback in terms of learner awareness, uptake, or more accurate subsequent performance with specifically targeted forms. Many of the studies have focused on verbal tense/aspect morphology or gender and number agreement, while others have focused on issues such as adverb placement, the position of clitics, or question formation. Common to all
of them is that the targets involve linguistic phenomena such as morphological markings or word orders which are represented phonologically, meaning that they cannot constitute tests of Carroll’s claims.

In studies of reactive feedback which have not set out with specific linguistic targets in view (e.g., Braidi, 2002; Carpenter et al., 2006; Egi, 2007a, 2007b, 2010; Ellis, Basturkmen, & Loewen, 2001; Gass & Lewis, 2007; Havranek, 2002; Heift, 2004, 2010; Kim & Han, 2007; Loewen, 2004, 2005; Loewen & Philp, 2006; Lyster, 1998a; Lyster & Mori, 2006; Lyster & Ranta, 1997; Mackey et al., 2000; Nassaji, 2007, 2009; Oliver, 2000; Oliver & Mackey, 2003; Panova & Lyster, 2002; Sheen, 2004, 2006; Williams, 2001; Yoshida, 2010; Zyzik & Polio, 2008), feedback episodes have often been classified as focusing on morphosyntax, phonology/pronunciation, or lexis/vocabulary. It is conceivable that some of these episodes may have focused on aspects of language not directly associated with some part of a phonological string, or with phonetic representations, but the authors have not mentioned these as important distinctions.
Table 1.1. Linguistic targets in studies on the effectiveness of feedback

<table>
<thead>
<tr>
<th>Study</th>
<th>Linguistic target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams (2007)</td>
<td>questions, locative prepositions, and past tense in English</td>
</tr>
<tr>
<td>Ammar &amp; Spada (2006); Ammar (2008)</td>
<td>third-person possessive determiners ‘his’ and ‘her’ in English</td>
</tr>
<tr>
<td>Ayoun (2001, 2004)</td>
<td>the passé composé and imparfait in French</td>
</tr>
<tr>
<td>Bigelow et al. (2006)</td>
<td>question forms in English</td>
</tr>
<tr>
<td>Bowles (2005)</td>
<td>dative experiencer constructions with the verb gustar in Spanish</td>
</tr>
<tr>
<td>Camblor-Portilla (2006)</td>
<td>noun-adjective gender and number agreement in Spanish</td>
</tr>
<tr>
<td>Carroll et al. (1992)</td>
<td>suffixation to form nouns in French</td>
</tr>
<tr>
<td>Carroll &amp; Swain (1993)</td>
<td>dative alternation in English</td>
</tr>
<tr>
<td>DeKeyser (1993)</td>
<td>regular adverb formation, use of the subjunctive after the verb vouloir (‘to want’), use of the imperfect in present counterfactual conditionals, reduction of the partitive article after adverbs of quantity, auxiliary use with the verb aller (‘to go’), and formation of il y a (‘there is/are’) in various tenses in French</td>
</tr>
<tr>
<td>Doughty &amp; Varela (1998)</td>
<td>simple past tense and past conditional sentences in English</td>
</tr>
<tr>
<td>Ellis (2007)</td>
<td>past tense (-ed) and comparatives in English</td>
</tr>
<tr>
<td>Ellis et al. (2006)</td>
<td>past tense (-ed) in English</td>
</tr>
<tr>
<td>Han (2002)</td>
<td>consistency of past tense marking in English</td>
</tr>
<tr>
<td>Herron &amp; Tomasello (1988)</td>
<td>direct object pronouns and negation in French</td>
</tr>
<tr>
<td>Ishida (2004)</td>
<td>aspect morphology (-te i-(ru)) in Japanese</td>
</tr>
<tr>
<td>Iwashita (2003)</td>
<td>vocabulary, a locative-initial sentence construction, and the –te form (an inflectional morpheme used to indicate progressive aspect) in Japanese</td>
</tr>
<tr>
<td>Lado (2008)</td>
<td>case morphology and subject-verb agreement as cues to semantic roles in Latin</td>
</tr>
</tbody>
</table>
Table 1.1. Linguistic targets in studies on the effectiveness of feedback (cont.)

<table>
<thead>
<tr>
<th>Study</th>
<th>Linguistic target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leeman (2003)</td>
<td>noun-adjective gender and number agreement in Spanish</td>
</tr>
<tr>
<td>Loewen &amp; Erlam (2006)</td>
<td>regular past tense in English</td>
</tr>
<tr>
<td>Loewen &amp; Nabei (2007)</td>
<td>question forms in English</td>
</tr>
<tr>
<td>Long et al. (1998)</td>
<td>adjective ordering and a locative construction in Japanese; object topicalization and adverb placement in Spanish</td>
</tr>
<tr>
<td>Lyster (2004); Lyster &amp; Izquierdo (2009)</td>
<td>grammatical gender in French</td>
</tr>
<tr>
<td>Mackey (2006)</td>
<td>question forms, plurals, past tense in English</td>
</tr>
<tr>
<td>Mackey (1999), Mackey &amp; Philp (1998); Mackey &amp; Oliver (2002); Mackey et al. (2002); Mackey &amp; Silver (2005)</td>
<td>question forms in English</td>
</tr>
<tr>
<td>McDonough (2005); McDonough &amp; Mackey (2006)</td>
<td>question forms in English</td>
</tr>
<tr>
<td>McDonough (2007)</td>
<td>past tense activity verbs in English</td>
</tr>
<tr>
<td>Moreno (2007)</td>
<td>assignment of semantic roles with pre-verbal direct object clitics in Spanish</td>
</tr>
<tr>
<td>Muranoi (2000)</td>
<td>indefinite articles in English</td>
</tr>
<tr>
<td>Nobuyoshi &amp; Ellis (1993)</td>
<td>past tense verbs in English</td>
</tr>
<tr>
<td>O’Relly, Flaitz, &amp; Kromrey (2001)</td>
<td>imperative forms in Spanish</td>
</tr>
<tr>
<td>Petersen (2010)</td>
<td>question forms in English</td>
</tr>
</tbody>
</table>
Table 1.1. Linguistic targets in studies on the effectiveness of feedback (cont.)

<table>
<thead>
<tr>
<th>Study</th>
<th>Linguistic target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philp (2003)</td>
<td>question forms in English</td>
</tr>
<tr>
<td>Révész (2009, to appear b); Révész &amp; Han (2006); Révész et al. (to appear)</td>
<td>past progressive verb forms in English</td>
</tr>
<tr>
<td>Rosa &amp; Leow (2004a,b)</td>
<td>past counterfactual conditional sentences in Spanish</td>
</tr>
<tr>
<td>Sachs &amp; Suh (2007)</td>
<td>backshifted verb tenses in indirect reported speech in English</td>
</tr>
<tr>
<td>Sagarra (2007)</td>
<td>noun-adjective gender and number agreement in Spanish</td>
</tr>
<tr>
<td>Sanz &amp; Morgan-Short (2004)</td>
<td>assignment of semantic roles with pre-verbal direct object clitics in Spanish</td>
</tr>
<tr>
<td>Sauro (2007, 2009)</td>
<td>zero article with generic non-count nouns in English</td>
</tr>
<tr>
<td>Sheen (2007, 2008)</td>
<td>articles in English</td>
</tr>
<tr>
<td>Tarone &amp; Bigelow (2007)</td>
<td>question forms in English</td>
</tr>
<tr>
<td></td>
<td>including comparative adjectives, placement of direct object pronouns, savoir versus connaître for ‘to know’</td>
</tr>
<tr>
<td>Trofimovich et al. (2007)</td>
<td>possessive determiners and lexis (verbs) in English</td>
</tr>
<tr>
<td>Yang &amp; Lyster (2010)</td>
<td>past tense forms in English</td>
</tr>
</tbody>
</table>

Probably correctly, Carroll (2001) states that (L2) “(d)evolutional research has remained the purview of researchers with no particular commitment to a theory of grammar” (p. 191). Another way of looking at this might be to state that researchers with such commitments do not tend to conduct developmental research. Either way, a great deal has been learned from L2 feedback studies without the selection of target structures’ having been
based on formal linguistic theory or ideas about the architecture of the language faculty. For researchers seeking to investigate learners’ noticing and learning processes in relation to different types of feedback, it has seemed natural to choose linguistic features based on what can be emphasized prosodically, bolded visually, referred to concretely, and so on. Carroll has a point, however: There is more to language than what is represented in surface forms, and it may be illuminating to investigate the effects of feedback on linguistic phenomena which depend on underlying structural relationships such as c-command, for example, and simply do not lend themselves to noticing in the sense of attending to morphosyntactic forms.

In order to conduct a full evaluation of Carroll’s predictions, it will ultimately be necessary to address the issue of awareness directly, and one way of approaching this will be to analyze introspective data from learners’ think-aloud protocols. As a first step, however, and as a preliminary test of Carroll’s claim that metalinguistic information should not be able to influence the learning of particular types of linguistic phenomena in principle, what this dissertation asks is whether feedback in a visual representational format (i.e., phrase-structure tree diagrams) can help to improve English-speaking learners’ accuracy of interpreting the reflexive *zibun* (‘self’) in L2 Japanese—a target involving structural constraints which are not phonologically represented and for which negative evidence should theoretically be necessary. At the same time, considering that L2 learners have differing profiles of cognitive abilities (e.g., in grammatical sensitivity, visuo-spatial memory, prior experience with metalinguistic concepts, and so on), some may be better predisposed to make use of metalinguistic visual feedback than others. This brings us to the issue of aptitude, the topic of Chapter 2, which will be followed by a discussion of the SLA of reflexives in Chapter 3.
(Re)conceptualizing aptitude

“Perhaps the clearest fact about SLA that we currently have,” according to Sawyer and Ranta (2001), is that adult L2 learners differ dramatically in both their rates of acquisition and in the ultimate attainment levels they eventually reach (p. 319). The fact, moreover, that individual differences (IDs) in language aptitude and motivation have been “the most consistent predictors of second language learning success” apart from age of onset poses what Dörnyei and Skehan (2003) have called a “correlational challenge” for the field of SLA (p. 589). Researchers are faced with the task of relating specific abilities to specific learning mechanisms. The existence of a variety of theories in the field means a wide variety of proposed mechanisms, from automatization processes in skill acquisition theory to cognitive comparisons in the interaction approach to parameter triggering in UG theories, and so on (Robinson, 2001, p. 378). Any of these might show associations (or dissociations) with cognitive and/or affective variables which shed greater light on the nature of adult SLA. From Skehan’s (2002) perspective, aptitude research will be “central for future progress in the second language learning domain” (p. 70).

As of 2001, however, when Sawyer and Ranta were writing, with the exception of Gardner’s work on motivation (e.g., Gardner & MacIntyre, 1992, 1993), “no major research program focusing on IDs and integrating them within a model of SLA [had] yet been developed” (p. 319). SLA research in the 1970s and 1980s tended to focus less on differences between learners and more on the exciting possibilities of universals and similarities among learners of
different ages and L1 backgrounds. Perhaps contributing to this, the notion of aptitude was considered by some to be “anti-egalitarian” and predictive of learning only in the context of “outmoded teaching methodologies” given that the standard test (J. B. Carroll and Sapon’s 1959 Modern Language Aptitude Test, or MLAT) had originally been validated against students’ grades in language courses employing audiolingual methods (Skehan, 2002, p. 72).

Although there had been researchers during those years who had argued for the importance of a wide range of individual-difference variables in SLA (e.g., intelligence, motivation, cognitive styles, strategies, anxiety, risk taking, etc.), Sawyer and Ranta were still able to claim in 2001 that research into aptitude-treatment interactions was “a huge and surprisingly uncharted area” (p. 352; see Spada, 2011, for more recent statements to the same effect). Sawyer and Ranta also suggested that early conceptualizations and methods of measuring aptitude had been perhaps more influential than might have been desired and may have “tended to blind later researchers to potentially more fruitful approaches” (p. 319). Skehan (2002) agreed: “Broadly speaking, [research conducted since the publication of the MLAT] has not reconceptualized aptitude in any significant manner, and most... has operated fairly clearly within the aptitude agenda set by Carroll” (p. 73). Of late, however, the field of SLA has taken some important initial steps toward updating theories and measurements of aptitude, and in suggesting avenues for future research.

The Modern Language Aptitude Test

The first and still most widely used test of L2 aptitude in SLA research, the Modern Language Aptitude Test (MLAT), was developed by John B. Carroll in the late 1950s, normed on
thousands of high school and college students as well as military personnel in 1958, and published in 1959 (J. B. Carroll & Sapon, 1959; for discussion, see J. B. Carroll, 1981, 1990; Stansfield & Reed, 2004; and Winke, 2005, 2011). Its purpose was to be able to select people who would be likely to progress quickly in language training; thus, it was designed to predict differences in the initial rate of instructed L2 learning in classroom contexts (J. B. Carroll, 1990; Robinson, 2005a; Skehan, 2002). The norming sample was taken from beginning language classes, and none of the learners had previously studied any other foreign languages. A major concern being feasibility of administration, the test development process involved administering a wide variety of different possible subtests (34), performing statistical analyses to determine which had the strongest predictive validity for course grades and the least overlap with other subtests, and considering which would be practical to administer. The final full version of the MLAT has five subtests, contains 149 items in total, and takes just over an hour to complete. Roehr and Gánem-Gutiérrez (2009, p. 166) point out that it has been endorsed as “one of the best instruments available” (e.g., by Sparks and Ganschow, 2001, among others), and Winke (2011) notes that, despite a variety of criticisms (to be discussed below), “[o]ver the decades, the predictive validity of the MLAT has prevailed” (p. 4).

In designing the MLAT’s subtests, J.B. Carroll tried to represent a variety of abilities he hypothesized to be involved in language learning. Ultimately, his model of L2 aptitude comprised four components: (1) phonetic coding ability (i.e., the ability to encode sounds and

---

3 The United States Government has continued to fund the development of other aptitude tests, such as the Defense Language Aptitude Battery (DLAB, Petersen & Al-Haik, 1976), VORD (Parry & Child, 1990), and the High-Level Language Aptitude Battery (Hi-LAB, Doughty et al., 2007, 2010), and tests have also been designed for younger populations, such as the Pimsleur Language Aptitude Battery (PLAB, Pimsleur, 1966) for students in grades 7-12. However, due in part to restrictions in availability of the former and the relatively young intended population of the latter, the MLAT has served as the “backbone” of aptitude testing and theorizing in adult SLA for the past 50 years (Fujii, 2005, p. 45).
to make and retain sound-symbol associations), (2) grammatical sensitivity (i.e., the ability to recognize the functions of linguistic entities in sentential contexts), (3) rote learning ability for linguistic material (i.e., the ability to form sound-meaning associations rapidly and then retain them), and (4) inductive language learning ability (i.e., the ability to infer the rules governing a rule-governed sample of linguistic data) (Carroll, 1981, p. 105). Carroll admitted, however, that the subtests on the MLAT are not precise operationalizations of exactly these constructs, and, in fact, inductive language learning ability is not measured directly, in part because the subtests which were designed to assess it were found to be less practical to administer.

In one sense, according to Skehan (1998), this lack of one-to-one correspondence may not be a problem; if language learning involves integrated sets of abilities, then a test drawing on integrated abilities can be serviceable (as discussed in Winke, 2005; see also Grigorenko, Sternberg, & Ehrman, 2000). Nevertheless, for SLA theory, as mentioned above, it is important for hypothesized aptitude components and measures to be situated within current approaches to language learning research and connected to particular SLA processes and mechanisms (Fujii, 2005). The constructs identified by Carroll were adduced prior to the advent of modern SLA theory, and many would argue that they represent neither a fine-grained enough nor a broad enough conception of L2 aptitude (e.g., Dörnyei, 2005; Sáfár & Kormos, 2008; Winke, 2011). As Winke (2011) puts it, however,

This does not mean that we should abandon the MLAT, but rather that the constellation of factors that contribute to successful language acquisition needs to be expanded beyond the constructs represented by the MLAT and more firmly situated within the context of learning (p. 6).
One way in which it has been suggested that conceptions of aptitude for L2 learning could be broadened is by including affective variables such as motivation, which is known to predict L2 learning success and has even been argued to be a necessary precondition for it (Csizér & Dörnyei, 2005; Winke, 2005, 2011). Another is by emphasizing learners’ creativity and ability to deal with novelty and ambiguity (Sternberg, 2002). Grigorenko et al. (2000) have accordingly developed the Cognitive Ability for Novelty in Acquisition of Language (Foreign) Test [CANAL-FT], which allows for dynamic simulation-based assessments of test-takers’ ability to learn at the time of testing. Test-takers are tasked with learning an artificial language, which is presented gradually over the course of the test in both auditory and visual formats, and they are evaluated on their immediate and delayed recall of information. This provides information on the learners’ strengths, weaknesses, and stylistic preferences, which can be used for diagnostic purposes to tailor instruction.

Fujii (2005) points out that the CANAL-FT is not yet widely recognized (p. 53), and its creators have acknowledged that it requires further validation; however, initial assessments of its content, construct, and criterion validity produced positive results (Grigorenko et al., 2000, p. 401). Furthermore, as will become clear below, several of the theoretical constructs underlying its development mesh with Skehan’s (2002) SLA-theory-based proposals regarding where new aptitude subtests might be valuable; for instance, the CANAL-FT was designed to test learners’ abilities to decide which information is relevant to attend to (selective encoding), to encode

---

4 The participants in the current study completed a series of motivation questionnaires over the course of the experiment; however, as those data have not yet been analyzed, this literature review will focus on cognitive aspects of L2 aptitude.
secondary information outside of their main attentional focus (accidental encoding), to draw on existing knowledge and determine the relevance of old information for current purposes (selective comparison), to apply rules in new contexts (selective transfer), and to synthesize disparate pieces of information and modify existing cognitive schemata (selective combination).

Focusing on these and other cognitive aspects of language learning ability, L2 aptitude can be characterized as “strengths individual learners have—relative to their population—in the cognitive abilities information processing draws on during L2 learning and performance in various contexts and at different stages” (Robinson, 2005a, p. 46). Skehan’s (2002) goals in a widely cited paper on reconceptualizing aptitude were to identify stages of L2 processing, link them to hypothesized aptitude components, and suggest directions for a theory-based research program in which new aptitude measures could be developed. Some of the initial stages he identified include noticing (i.e., attending to particular linguistic elements), which is hypothesized to require aptitude in auditory segmentation, attention management, working memory, and phonemic coding; and pattern identification (i.e., making implicit or explicit generalizations based on perceived regularities), which is hypothesized to require working memory and grammatical sensitivity. For later stages of extending and complexifying (i.e., widening the domain of a generalization and restructuring it in light of new information), grammatical sensitivity and inductive language learning ability are argued to be important. Then, in order for learners to become more accurate, create an accessible repertoire, automatize rule-based language, and lexicalize to achieve greater fluency, Skehan suggests the importance of proceduralization, automatization, chunking, and retrieval processes (pp. 88-92).
For some of these components (e.g., working memory, phonemic coding, grammatical sensitivity), preliminary measures, at least, exist, whereas in other cases (e.g., the ability to deal with ambiguity) new tests may need to be designed. In the present study, for which the learning target was interpretations of the reflexive *zibun* (‘self’) in Japanese, it seemed potentially worthwhile to gather information on learners’ sensitivity to linguistic ambiguity. As will be reviewed in Chapters 3 and 4, in both English and Japanese, sentences containing reflexives can be ambiguous. Since previous research (e.g., MacLaughlin, 1998; Thomas, 1991) has indicated that people differ in their sensitivity to multiple possible interpretations of reflexives even in their native languages, a test (described in Chapter 4) was designed to tap this ability as an individual-difference variable that might have an impact on learning in the context of this experiment.

In a more general sense, perhaps the most prominent development in L2 aptitude research has been (and will be) in the area of memory, which, according to Skehan (2002), “is the aspect of the MLAT which has least stood the test of time” (p. 71). In contrast to earlier conceptions of memory as fairly static, passive, and associational in nature, current work in cognitive psychology highlights more dynamic properties, such as the use of executive control to allocate attentional resources to both storage and processing.

**Working memory**

When processing sentences in either an L1 or an L2, it is necessary to relate and integrate discontinuous parts of a superficially linear stream of language. This requires the ability to store linguistic elements temporarily while simultaneously continuing to process
incoming information and while inhibiting potential distractions which might interfere. In other words, it requires working memory (WM) capacity, generally defined as “the ability to maintain information in an active and readily accessible state, while concurrently and selectively processing new information” (Conway, Jarrold, Kane, Miyake, & Towse, 2007, p. 3). The cognitive resources supporting this maintenance and manipulation of information are dissociable and “flexibly deployable,” according to Miyake and Friedman (1998), but also limited and subject to inter-individual variation (p. 341).

The cognitive psychology literature contains an extensive array of theoretical approaches to WM, many of them in partial opposition to each other (for overviews, see, e.g., Baddeley, 2007; Conway et al., 2007; Shah & Miyake, 1999). Fortunately, Goo (2010, pp. 717-20) provides a clear and concise overview for SLA researchers, briefly describing Daneman and Carpenter’s (1980, 1983) resource-sharing account (in which WM is construed as involving trade-offs between processing and storage); Towse and Hitch’s (1995, 2007) task-switching account (which focuses on processing efficiency and the temporal dynamics of information maintenance, with limitations in WM seen as being related to the rapid fading of memory traces during processing); Conway, Engle, Kane, Tuholski, and colleagues’ executive attention account (which highlights variation in both domain-general executive functions and domain-specific storage and rehearsal; e.g., Engle, 2002; Engle, Kane, & Tuholski, 1999; Kane, Conway, Hambrick, & Engle, 2007); Hasher, Zacks, and colleagues’ inhibition-based account (which emphasizes cognitive control processes that inhibit and delete irrelevant information and restrain ‘prepotent’ responses so that weaker ones can be evaluated in light of current goals; e.g., Hasher, Lustig, & Zacks, 2007); and Jarrold and Bayliss’ (2007) coordination-oriented
account (which stresses multiple independent sources of variation, including storage constraints, processing efficiency, and executive control and coordination ability).

Although there are important differences among these theoretical approaches, what many researchers in the field of SLA seem to agree on is that since individual differences in WM can affect the efficiency and quality of L2 processing in real time, working memory may be a “central component of language aptitude” (a frequently cited proposal made by Miyake and Friedman, 1998, p. 340; see also Robinson, 2005a; Skehan, 2002). Experimental methods in this emerging area of SLA research are still rather unstandardized among L2 researchers, and results have been somewhat mixed (Juffs, 2004; Winke, 2011); however, several empirical studies, conducted with L2 learners of a variety of ages, suggest that various components of WM can play a role in a variety of aspects of SLA, including vocabulary learning (e.g., Atkins & Baddeley, 1998; Cheung, 1996; Daneman & Case, 1981; Dufva & Voeten, 1999; Hu, 2003; Hummel, 2009; Masoura & Gathercole, 1999; Papagno, Valentine, & Baddeley, 1991; Papagno & Vallar, 1992, 1995; Service, 1992; Service & Craik, 1993; Service & Kohonen, 1995; Speciale, N. Ellis, & Bywater, 2004) and grammar learning (e.g., N. Ellis & Sinclair, 1996; French & O’Brien, 2008; Goo, 2010; Hummel, 2009; Kempe & Brooks, 2008; Lado, 2008; Leeser, 2007; Mackey & Sachs, to appear; O’Brien, Segalowitz, Collentine, & Freed, 2006; Révész, to appear (b); Robinson, 2002, 2005b; Trofimovich et al., 2007; Williams, 1999; Williams & Lovatt, 2003); as well as in interactional processes such as learners’ noticing of feedback (Mackey et al., 2002) and production of modified output (Mackey et al., 2010; Sagarra, 2007a); in the relative benefits of different instructional methods (e.g., Ando, Fukunaga, Kurahashi, Suto, Nakano, & Kage, 1992; Erlam, 2005); and in L2 oral fluency (e.g., O’Brien, Segalowitz, Freed, & Collentine, 2007).
Working memory also appears to play a role in L2 sentence processing, influencing learners’ preferences for word-order versus animacy and case-marking cues (e.g., Miyake & Friedman, 1998), their online processing of redundant morphology (e.g., Sagarra, 2007b), and their resolution of syntactic ambiguities (e.g., Havik, Roberts, van Hout, Schreuder, & Haverkort, 2009); and it has been associated with L2 comprehension, production, and proficiency more generally (e.g., Abu-Rabia, 2003; Geva & Ryan, 1993; Harrington & Sawyer, 1992; Kormos & Sáfár, 2008; Leeser, 2007; Payne & Ross, 2005; Payne & Whitney, 2002; Walter, 2004). Inspired by findings such as these, increasing numbers of SLA researchers are including WM measures in their studies to supplement traditional measures of L2 aptitude.

In Baddeley’s (2003, 2007) well-known model, often referred to in SLA studies, the working memory system is argued to comprise multiple components, including (at least) a central executive involved in allocating attention and regulating information processing, an ‘episodic buffer’ to integrate information from different modalities, and 2 ‘slave systems’ for the storage of modality-specific information: a phonological loop and a visuo-spatial sketchpad. Many of the studies listed above (e.g., Mackey et al., 2010; Sagarra, 2007a) have focused on participants’ executive ability to store and process information simultaneously, whereas others (e.g., Speciale et al., 2004) have focused on their ability to recall strings of phonological input. Still others have measured both (e.g., Kormos & Sáfár, 2008; Mackey & Sachs, to appear; Révész, to appear (b)). Baddeley (2003) points out, however, that the relationship between visuo-spatial memory and language learning remains relatively unexplored. Considering the status of phonology as one of the major subsystems of language, a focus on the role of phonological loop capacity in SLA makes sense. That being said, above and beyond the requirements of
developing literacy skills in a new orthography, for example (which, it would seem, must involve visual memory), there are also linguistic phenomena (e.g., locative case particles or adpositions, manner-of-motion verbs) which might be hypothesized to involve a visuo-spatial orientation. Furthermore, in L2 pedagogy, there may be instructional methods that draw on visual memory.

In the present study, one of the treatment conditions was purposely designed to be visual in nature, making use of tree diagrams to illustrate the structural relationships between reflexives and their antecedents. Thus, it was deemed important to include a measure of visual short-term memory (STM), the Visual Patterns Test (Della Sala, Gray, Baddeley, & Wilson, 1997), described in more detail in Chapter 4.5

**Familiarity with metalinguistic concepts and terminology**

As we have seen in Chapter 1, many researchers (e.g., N. Ellis, 1994, 2005; R. Ellis, 1993, 1994, 2004, 2009; Schmidt, 1994, 2001; Sharwood Smith, 1991) have argued for a ‘weak interface’ between explicit and implicit learning processes whereby explicit information about a language can influence learners’ attention to input and production of output, thereby potentially influencing the parallel development of implicit knowledge. With this as a mindset, it seems natural to hypothesize that learners who have richer explicit knowledge, and/or a greater capacity to apply their explicit knowledge, might be better positioned to profit from this means of learning. More accurate metalinguistic concepts, brought to bear on input or output

---

5 Two measures of executive control were also administered: (1) a listening-span test based on Daneman and Carpenter’s (1980) reading-span task, used with the permission of Cristina Sanz, Harriet Wood Bowden, and Catherine Stafford, who developed it with the assistance of Bill Garr for their research in *The Latin Project*; and (2) an operation-span test based on a task originally developed by Turner and Engle (1989), and modified from a version provided by Jaemyung Goo, a colleague at Georgetown University. However, as those data have not yet been analyzed, the Methods chapter focuses on the administration of the Visual Patterns Test.
at opportune times in appropriate ways, might be able to shape L2 processing and production more beneficially.

Indeed, it has frequently been found that participants in L2 learning experiments who report awareness of linguistic targets at the level of understanding (Schmidt, 1990, 2001), showing conceptually driven processing, demonstrate greater L2 development than participants who report awareness at the lower level of noticing or report no awareness at all (e.g., Leow, 1997; Rosa & O’Neill, 1999; Rosa & O’Neill, 2004a). Although causality cannot be inferred from these results (since certain abilities, knowledge, and “attentional predispositions” [Birdsong, 1989] may cluster together), they are at least consistent with the ideas above, as well as with Gass and Selinker’s (2001) observation that the “ability to think about language is often associated with an increased ability to learn a language” (p. 302). If there are, in fact, circumstances in which awareness at the level of understanding is necessary for L2 learners to overcome the automatic processing biases of their L1s, as Schmidt (2001) speculates, and if metalinguistic knowledge can facilitate the experiencing of awareness at the level of understanding, then one might expect to find that learners with greater explicit knowledge of language show greater improvement under those circumstances.

An important consideration here is the fact that the external provision of metalinguistic information (e.g., via feedback or prior instruction) is not always found to be effective, or is not always found to be more effective than more implicit instructional treatments (e.g., Moreno, 2007; Sanz & Morgan-Short, 2004; VanPatten & Oikkenon, 1996). In relation to this, researchers (e.g., DeKeyser, 2005; Lado, 2008) have suggested as possible mediating factors the complexity of the target forms and the learners’ prior knowledge and cognitive abilities, likely in interaction
with one another. These relationships may play out in multiple ways; for instance, hypothetically, for relatively simple linguistic targets, if the provision of metalinguistic information benefits anyone, it may be likely to benefit learners with high enough grammatical sensitivity to make use of it, but less pre-existing metalinguistic knowledge (assuming that learners with more metalinguistic knowledge might not need the additional information). For complex linguistic targets, on the other hand, if metalinguistic information benefits anyone, it may be likely to benefit learners with high enough grammatical sensitivity and more pre-existing metalinguistic knowledge (because they might be better prepared to understand it). To investigate the possibility of intricate relationships such as these, it is important to develop clear definitions of various types of metalinguistic abilities, as well as fine-grained instruments for measuring them.

Combining aspects of definitions from R. Ellis (2004, 2009), Elder (2009), and Purpura (2004), metalinguistic knowledge can be defined as a set of informational structures, available for use from long-term memory, which have to do with the attributes of language, including issues of phonology, morphology, syntax, semantics, and pragmatics. It can be described as analytical in nature and at least potentially verbalizable, involving “explicit declarative facts (whether rules or fragments of information)” (Elder, 2009, p. 114). Taking a metalinguistic approach involves stepping back from language as a transparent medium of communication and treating it instead as an object that can be analyzed (Gass & Selinker, 2001, p. 302). Learners tend to know when they are drawing on metalinguistic knowledge (Elder, 2009, p. 114) and can use it as “a ‘tool’... to mediate performance and achieve self-control in linguistically demanding situations” (R. Ellis, 2009, p. 13); however, it may not always be retrievable in real
time on tasks that involve making use of automatized knowledge under time pressure. The prototypicality of the linguistic phenomena in question can also affect learners’ ability to apply their metalinguistic knowledge in a given situation (Hu, 2002).

In reasoning about the potential contributions of metalinguistic knowledge to aptitude for L2 learning, it is important to recognize at least two facts: (1) that metalinguistic knowledge, and even the verbal expression of it, does not necessarily require the possession of formal metalinguistic terminology; and (2) that learners’ metalinguistic rules are quite often incomplete, imprecise, inaccurate, and even anomalous (Y. Han & R. Ellis, 1998; Sorace, 1985).

Regarding the first point, R. Ellis (2004) draws a distinction between what he calls ‘analyzed’ or ‘explicit’ knowledge (i.e., conscious representations of linguistic structures) and ‘metalinguistic’ knowledge (i.e., the possession of labels or terminology). Other authors use the terms differently, but make a similar distinction; for instance, Berry (2009) points out that what he calls ‘metalinguistic’ knowledge (i.e., knowledge about language) is not the same thing as ‘metalingual’ knowledge (i.e., knowledge of metalanguage). In Elder’s (2009) view, although a command of metalinguistic terminology might help a person to display his or her knowledge, the knowledge of “subject-specific lexis (verb, noun, etc.)... is independent of grammatical knowledge per se... and indeed of any cognitive or analytical skills associated with such knowledge” (pp. 114-115). Drawing an analogy with architecture, she points out that one can know how to design a building without necessarily being able to label all of the parts. As far as usefulness for SLA is concerned, R. Ellis (2004) maintains that it is probably “learners’ understanding of explicit linguistic constructs rather than their ability to articulate metalinguistic rules that is important” (p. 267); nonetheless, he does note that the two “would
seem to be closely related” (2009, p. 13), and he contends that “metalinguistic knowledge [i.e.,
of terminology] may assist learners in developing explicit knowledge that has greater precision
and accuracy” (2004, p. 261). When learners have access to technical metalinguistic terms, their
knowledge may be “clearer and better structured” (Y. Han & R. Ellis, 1998, p. 6).

Regarding the second point, that learners’ knowledge of grammar is often vague or
inaccurate, Elder (2009) states that this can improve over time, along with increases in L2
proficiency and in relation to the type of instruction received. For instance, Elder and
Manwaring (2004) found that learners of Chinese who had received just one year of intensive
grammar-oriented university-level L2 instruction demonstrated greater metalinguistic
knowledge than fellow students who had studied Chinese for several years in more
communicative classes in secondary school. A common finding in studies testing university
students’ metalinguistic knowledge has been that there is wide variation in their familiarity with
terms for grammatical constructs (e.g., Alderson, Clapham, & Steel, 1997; Berry, 2009; Green &
Hecht, 1992). Alderson et al. (1997), for instance, gave 509 British university students a test
which involved identifying various parts of speech in English (L1) and French (L2) sentences,
then correcting errors and stating rules relevant to the corrections of other sentences. Scores
ranged from 14% through 100%, with an average of about 49%, and the authors note that
although most students seemed to be familiar with a few basic terms, such as noun and verb,
many other terms, such as indefinite article and finite verb, were unfamiliar to more than half of
the students (p. 106). Similarly, in a study with L2 English majors at universities in Austria,
Poland, and Hong Kong, Berry (2009) found variation both within and across countries in
students’ knowledge of metalinguistic terminology. His test presented the participants with a
list of 50 grammatical terms and asked them either to provide a stand-alone example of each or
to underline relevant word(s) in a phrase or clause which they themselves had written. As in
Alderson et al.’s (1997) study, the most commonly known terms included basic parts of speech,
such as noun, verb, and adjective, along with certain words related to morphology, such as
plural. However, terms which Berry deemed more ‘scientific’ (as opposed to pedagogical) in
nature, such as complement and finite verb, were much less well known, and others (e.g.,
definite article, conjunction) showed tendencies to be well known in one country while being
relatively unknown in another.

In the studies which have been conducted to investigate relationships between learners’
metalinguistic knowledge on the one hand and their language aptitude and/or L2 proficiency on
the other (e.g., Alderson et al., 1997; Berry, 2009; Elder, 2009; Elder & Manwaring, 2004; Y. Han
& R. Ellis, 1998; Roehr, 2006, 2008; Roehr & Gánem-Gutiérrez, 2009), the operationalizations of
metalinguistic knowledge and L2 proficiency have varied, and so have the results. In Alderson et
al.’s (1997) study, in addition to the test of metalinguistic knowledge described above, the
researchers administered MLAT Part IV (Words in Sentences, considered to measure
grammatical sensitivity), a Swahili-based test of inductive language learning ability, a multiple-
choice grammar test, and tests of general L2 proficiency, reading comprehension, listening, and
writing. Regarding the question of a relationship between metalinguistic knowledge and L2
(French) proficiency, they report finding “no evidence to support the belief that students with
higher levels of metalinguistic knowledge perform better at French, or that they improve their
French proficiency at higher rates than other students” (p. 118). Regarding a potential
relationship with language aptitude, they acknowledge moderate correlations between the
MLAT on the one hand and both metalinguistic knowledge and L2 proficiency on the other (p. 116); however, finding that MLAT scores did not load with either in a factor analysis (or indeed on any factor), they state that “[t]he contribution of aptitude... to metalinguistic knowledge and language proficiency is ambiguous” (p. 118).

In a study by Y. Han and R. Ellis (1998), focusing on advanced university ESL learners’ knowledge of verb complementation (i.e., whether particular verbs subcategorize for gerunds, infinitives, finite clauses, etc.), the researchers likewise did not find a relationship between the learners’ metalinguistic knowledge of rules and their L2 proficiency; however, they note that verb complementation seems likely to “[lend] itself to item- rather than rule-learning” (p. 16), and given “the absence of a clear rule to guide learning” in that area, they consider it unsurprising that the participants had difficulty verbalizing knowledge about it (p. 17).

In contrast, several other studies have uncovered relationships among metalinguistic knowledge, aptitude, and L2 ability. Berry (2009), for example, found an association between Hong Kong students’ familiarity with metalinguistic terms and their L2 proficiency.6 Roehr and Gánem-Gutiérrez (2009) investigated the relationships among L2 learners’ metalinguistic knowledge, language aptitude (as measured by the MLAT), and working memory. The participants were 39 L1-English university students, 19 of whom were studying L2 German, and 20 of whom were studying L2 Spanish. The tests of metalinguistic knowledge were given in the learners’ L2s and included 2 sections: one which involved correcting, describing, and explaining errors in L2 sentences, and one (modeled on the MLAT, Part IV) which involved recognizing the grammatical roles of words in L2 sentences without having to label, describe, or explain them.

---

6 The Hong Kong group was the only one which could be included in that analysis since only they had reported scores on a pre-university exam of English ability.
Roehr and Gánem-Gutiérrez found a weak to moderate positive correlation between the learners’ L2 metalinguistic knowledge and their performance on the MLAT as a whole ($r = .34$). Unsurprisingly (considering that MLAT IV had served as a model for part of the metalinguistic knowledge test), the correlations with MLAT IV ($r = .41-.45$) were stronger than with any other section of the MLAT ($r = .06-.36$). In a principal components analysis, however, metalinguistic knowledge emerged as its own component, separate from one including MLAT Parts I (Number Learning), II (Phonetic Script), and IV (Words in Sentences); another including the working memory tests; and another including MLAT Parts III (Spelling Clues) and V (Paired Associates). The authors interpret this as indicating that metalinguistic knowledge is distinguishable from language aptitude, and even from that aspect of aptitude which involves identifying the roles of words in sentences (pp. 174-5), though it seems very important to keep the nature of the operationalizations in mind: The MLAT was conducted in the participants’ L1, with certain sections being “highly speeded” and difficult to complete in the time provided (Carroll & Sapon, 2002, p. 3), whereas the metalinguistic knowledge test was untimed and involved analyzing L2 sentences, which seems likely to depend also on L2 proficiency.

Elder (2009) reports on a study conducted to evaluate the validity of a test of metalinguistic knowledge with regard to several aspects of the definition reviewed above (e.g., that it is explicit and declarative, less accessible under time pressure, etc.). The participants were 20 native speakers of English and 229 speakers of other L1s who had been learning L2 English for an average of 10 years. In addition to the test of metalinguistic knowledge, they were administered tests of explicit and implicit English knowledge (untimed and timed grammaticality judgment tasks [GJTs], an oral narrative task, and an elicited oral imitation test,
all focusing on the same linguistic targets). Data were also gathered on the participants’ self-assessed levels of grammatical knowledge and on their L2 proficiency (as measured by the TOEFL, IELTS, and/or the Diagnostic English Language Needs Assessment, a test of academic language skills). The metalinguistic knowledge test contained two parts: one which involved identifying examples of grammar terms in the context of an L2 sentence, and another which involved choosing from sets of multiple-choice options the appropriate pedagogical rules to explain highlighted errors in ungrammatical L2 sentences. The former was intended to measure familiarity with metalinguistic terminology, whereas the latter was intended to measure explicit knowledge of the L2. Elder notes, however, that most of the options in the latter section also contained “basic metalinguistic terminology” (e.g., conditional, subjunctive, past perfect) because “it is often difficult to formulate explanations without the use of such terms” (p. 118). She points out as well that the test items “varied considerably... in the amount of technical terminology used and also in the extent to which this terminology was crucial for understanding the rule” (p. 134). Thus, knowledge of metalinguistic terms and L2 proficiency seem likely to have been implicated in both sections of the test.

The study produced strong evidence that the metalinguistic knowledge test tapped explicit/declarative knowledge which was distinct from implicit/proceduralized knowledge. A confirmatory factor analysis indicated that a 2-factor solution was a good fit to the data, with the elicited imitation task and timed GJT loading on one factor (implicit knowledge) and the metalinguistic knowledge test and ungrammatical sentences of the untimed GJT loading on the other (explicit knowledge). There was a weak but statistically significant correlation between the participants’ self-reported use of rules on the untimed GJT and their performance on the
rule-explanation section of the metalinguistic knowledge test. There was also a moderate and significant positive correlation between the learners’ self-assessed levels of grammar knowledge and their scores on the test of metalinguistic knowledge, whereas this was not the case for the tests of implicit knowledge. Metalinguistic knowledge was not associated with accuracy of performance on the timed oral narrative task, and the correlations between metalinguistic knowledge and reading proficiency were consistently stronger than those between metalinguistic knowledge and listening skills, where time pressure may afford less opportunity for controlled processing. Each of these results provided either partial or full support for the study’s initial hypotheses.

One finding which ran contrary to what was expected was that of a relatively strong positive relationship ($r = .55^*$) between scores on the different sections of the metalinguistic knowledge test. Elder points out that learners sometimes did demonstrate explicit knowledge of specific L2 rules without showing a command of the relevant metalinguistic terminology as applied to examples in the L2, and vice versa. However, in relation to this, and also in relation to the designs and results of the other studies reviewed above, where operationalizations of metalinguistic knowledge have often seemed tied to knowledge of the L2 specifically, it may be worthwhile to consider Bialystok’s (2001) recommendations regarding uses of the term ‘metalinguistic’. She argues that conceptions of metalinguistic knowledge should involve a “higher level of generality than... the specific details of any particular language” (p. 127)—that is, they should involve abstract knowledge about language, as opposed to explicit knowledge of L2 grammar. In the present study, as described in Chapter 4, the test of familiarity with metalinguistic terminology was given entirely in English, the participants’ dominant language.
Over the decades, there has been a persistent notion in the field of SLA that aptitude, as measured by the MLAT, might be relevant only in instructed settings. One reason for this derives straightforwardly from the context in which the MLAT was developed (i.e., focusing on predictive validity for performance in language courses), but from a theoretical perspective also, it has been argued that the ability to analyze language should affect only ‘learning’ (of explicit facts about language, which can be used for conscious monitoring), and not ‘acquisition’ (of implicit competence) (e.g., Krashen, 1981, 1982, 1985). Some early research, conducted under an artificial grammar paradigm and using the Wechsler Adult Intelligence Scale to measure IQ, did find that participants showed weaker relationships between intelligence and performance under implicit learning conditions (e.g., Reber, Walkenfeld, & Hernstadt, 1991). However, several SLA studies have found verbal analytic ability to be related to L2 development under experimental conditions both with and without explicit rule presentation (e.g., de Graaff, 1997; Robinson, 1997b), as well as in communicative and immersion classroom contexts (e.g., Ehrman & Oxford, 1995; Fujii, 2005; Harley & Hart, 1997; Ranta, 2002).

One approach to explaining this would assemble arguments from a weak-interface perspective, as discussed above. Another, by Robinson (1997b), would propose a so-called “fundamental similarity” in adult SLA processes under any condition of L2 exposure. Robinson reasons that learning a second language as an adult necessarily “draws on abilities implicated in explicitly processing input, and ‘noticing’ features of it” (2005a, p. 63). Likewise, Skehan (2002) claims that the processing stages and aptitude components he has outlined should be relevant “not simply for formal classroom learning, but for any SLA” (p. 69). In Skehan’s (1998) view, it
makes excellent sense that grammatical sensitivity would be relevant in communicative contexts. In fact, perhaps particularly in situations without as much instructional scaffolding, it may be analysis-oriented learners with higher grammatical sensitivity who are better able to focus their attention efficiently, notice relevant input, and impose structure on it. J. B. Carroll, the developer of the MLAT, similarly suggested that high-quality instruction might make certain kinds of aptitude less decisive (1963, discussed in Sawyer & Ranta, 2001, p. 327).

Importantly, proposing that adult SLA is “fundamentally similar” under any condition of L2 exposure does not amount to a claim that all abilities should be equally influential across all contexts. The idea that different learning conditions can draw more or less heavily on different profiles of individual characteristics has been researched extensively in the fields of education and psychology (see, e.g., Cronbach & Snow, 1977; Snow, 1989, 1991) and has led to the conclusion that aptitude is a “person-in-situation transaction” (Robinson, 2005a, p. 61, discussing the ideas of Richard Snow). According to Snow (1987), aptitude is best understood as “an interface between an inner environment... and an outer environment in which it operates” (p. 12), and “describing the situation is part of defining the aptitude” (p. 13, emphasis in the original, cited by Fujii, 2005, p. 227). As Robinson (2005a) puts it, “learning contexts..., the pedagogic interventions taking place within them..., and the cognitive processes they implicate... all have the effects they do in interaction with the patterns of abilities learners bring to those contexts” (p. 47), and learning might be optimized by matching instructional strategies to learners’ individual profiles.

Recently, SLA research along these lines has been gaining momentum, with at least two major goals—a theoretical one being to reach a greater understanding of language learning
processes by discovering which abilities contribute to success under what circumstances, and a pedagogical one being to attain a greater ability to tailor teaching methods to learners’ aptitudes. In order to do this, Snow (1994) argues, “the relevant aspects of person and situation [must be] specified, their interaction [must be] demonstrated empirically, and some process explanation of how and why this occurs [must be] offered” (p. 4, cited in Robinson, 2001, p. 370). In the field of SLA, the construct of ‘aptitude’ is no longer viewed simply as a predictor of language learning success in general; studying aptitude necessarily means studying aptitude-treatment interactions (ATI) (Fujii, 2005, p. 70).

**Aptitude-treatment interactions**

Three decades ago, Wesche (1981) was already exploring how L2 learners (in this case, Canadian civil servants in a language training program) might be matched to instructional methods that fit their profiles of abilities. Placement testing for the program involved taking the MLAT and parts of the PLAB, and the learners’ outcomes resulted in their being classified as either analysis-oriented (with high grammatical sensitivity, but possibly low memory/auditory abilities), memory-oriented (with high memory/auditory abilities, but possibly low grammatical sensitivity), or neither (‘flat’ profiles with no clear strengths or weaknesses). The students with identifiable orientations were assigned either to courses focusing on grammatical analysis via abundant written input or to courses focusing on communication in interpersonal situations. The rest were assigned to the default course type, using audiolingual methods. What Wesche found was that learners who had been placed in courses that matched their profiles not only reported greater satisfaction but also tended to show higher levels of L2 achievement.
Based on these findings, as well as on his own research (Skehan, 1986) using cluster analysis to identify different ability profiles among British army personnel in an intensive Arabic course, Skehan (1998) proposed that L2 learners may differ in terms of “predisposition[s] to view language learning as syntactic or lexical” (1998, p. 193). Whereas learners strong in analytic skills might focus on finding patterns and developing a system of rules to deploy, learners strong in memory skills might focus on remembering chunks and developing a repertoire of lexicalized strings. More recently, Hummel (2009) has suggested that individuals can actually “build their own idiosyncratic patterns of abilities in L2 learning” (p. 241) and that a variety of ability profiles can lead to successful L2 performance. In her study, phonological memory and “traditional aptitude components” involving the ability to detect linguistic patterns were not found to be related to each other, but each predicted L2 proficiency among French speakers enrolled in a university program in Teaching English as a Second Language.

Wesche and Skehan’s early research focused on two broad combinations of aptitude components: high memory/low analysis and low memory/high analysis (Sawyer & Ranta, 2001, p. 352). Taking a finer-grained view nowadays, as discussed above, Skehan (2002) has put forward a wide range of aptitude components to correspond with different stages of L2 processing and SLA, and Robinson (2001, 2005a) has emphasized, in his Aptitude Complexes Hypothesis, that combinations (or ‘complexes’) of abilities “jointly facilitate processing and learning…. from focus on form techniques that draw on them” (2005a, pp. 51, 63) “under different psycholinguistic processing conditions” in specific instructional contexts (2001, p. 329). In other words, according to Robinson, it is even more intricate than specific abilities being associated with specific stages: Particular sets of cognitive resources (e.g., processing speed,
attention management, pattern recognition) underlie primary abilities (e.g., working memory, grammatical sensitivity) which, in different combinations, form higher-order abilities (e.g., memory for contingent speech, noticing the gap, metalinguistic rule rehearsal) which, also in different combinations, can put learners in a position to gain greater benefits at various stages of L2 processing under particular circumstances. Robinson (2005a, p. 51) has proposed, for instance, that memory for contingent speech and noticing the gap (each of which is made up of more basic abilities) combine to form an aptitude complex that facilitates learning from recasts.

Several empirical studies have investigated aptitude-treatment interactions in both classroom- (e.g., DeKeyser, 1993; Erlam, 2005; Fujii, 2005; Sheen, 2008) and laboratory-based contexts (e.g., de Graaff, 1997; Lado, 2008; Révész, to appear (b); Robinson, 1997b, 2002). The classroom-based research has revealed a variety of relationships among learner characteristics, features of instructional conditions, and aspects of L2 performance, while also highlighting the fact that higher-order interactions between variables can sometimes obscure the links one might otherwise expect to find based on both theory and previous empirical studies. As one fabricated but not implausible example, for the sake of illustration, a researcher might find language analytic ability to be associated with the extent to which learners benefit from oral feedback, but only for low-anxiety learners with high working memory who have a tendency to devote attentional resources to focusing on form because they are not particularly concerned with the most efficient possible completion of a communicative task—or the researcher might not find any of this because too many variables are involved. As another example, inspired by an actual study by DeKeyser (1993) to be discussed below, learners’ degrees of extrinsic motivation might unexpectedly turn out to be negatively related to the extent to which they
benefit from externally-provided feedback, but this might seem explicable in light of a strong positive relationship between extrinsic motivation and anxiety, together with a strong negative relationship between anxiety and benefiting from feedback. Such relationships make sense, but the more of them there are, the more difficult it can be for researchers to identify them.

Classroom-based studies of aptitude-treatment interactions

In DeKeyser’s (1993) study of 35 Dutch-speaking high-school learners of L2 French, he set out to investigate how cognitive and affective variables might be related to whether corrective feedback had a positive or negative impact in a classroom setting. The participants were students in two intact classes, one of whose teachers was asked to provide error correction “as frequently and explicitly as possible” (p. 505). The other teacher was asked to avoid correcting students’ mistakes. DeKeyser measured the students’ grammatical sensitivity using a Dutch version of MLAT Part IV and measured their extrinsic motivation and anxiety using Likert-style questionnaire items. He also assessed the participants’ French proficiency (using 3 communicative tasks performed with a native speaker of French) and grammatical achievement (using a 60-item fill-in-the-blank test focusing on 6 areas of French grammar) both before and after the year-long treatment.

Among DeKeyser’s findings were that students with high pre-test scores and low levels of anxiety showed greater benefits from error correction, as measured by the written grammar test, and that whereas students with low extrinsic motivation showed greater oral accuracy and fluency when they had been systematically corrected, students with high extrinsic motivation showed greater oral accuracy and fluency when they had not been corrected. The direction of
this relationship was the reverse of what he had hypothesized, in response to which he 
reasoned that it was “probably due to the correlation [.52] between motivation and anxiety” in 
the error-correction group (p. 510). Contrary to another of his hypotheses, DeKeyser did not 
find a relationship between the students’ grammatical sensitivity and the amount they 
appeared to have benefited from error correction, but he suggested that this might similarly be 
attributed to an interaction with anxiety (that is, the interaction between aptitude and error 
correction might have been washed out or obscured by a stronger relationship with anxiety). 
Thus, he interpreted his results as “evidence for the hypothesis of interaction between error 
correction and anxiety rather than as evidence against the hypothesis of interaction between 
error correction and aptitude” (p. 511). DeKeyser concluded, notably, that although error 
correction did not seem to show an effect on learning overall, it did show clear effects for 
students with either very high or very low scores on particular ID measures (i.e., previous 
achievement, extrinsic motivation, and anxiety).

Another study investigating the effectiveness of feedback in relation to learner anxiety 
in intact classes was carried out by Sheen (2008) with 45 intermediate-level ESL learners from a 
variety of L1 backgrounds. The linguistic target was two functions of articles: the use of a when 
a referent is first mentioned, and the use of the for anaphoric reference. Her research questions 
asked whether there were relationships between (1) L2-classroom anxiety and the effects of 
recasts on learners’ grammatical accuracy and (2) L2-classroom anxiety and learners’ responses 
to recasts in terms of modified output and repair. Based on the participants’ responses on an 8-
item Likert-style questionnaire, Sheen divided them into 4 groups: 13 high-anxiety learners and 
11 low-anxiety learners who received recasts, and 11 high-anxiety learners and 10 low-anxiety
learners who did not receive recasts. (She excluded from these groups any learners whose anxiety scores were within 1 standard deviation of the mean.) All of the participants completed pre-tests, immediate post-tests, and 4-week delayed post-tests, in 3 forms each: a 14-item speeded dictation task intended to measure implicit knowledge, a 17-item written error-correction task, and a picture-based narrative-writing task. The participants in the non-recast (control) groups completed only the tests, not engaging in any particular treatments beyond their normal attendance in class. The participants in the recast groups, on the other hand, engaged in two classroom-based treatment sessions which involved reconstructing a story in groups of 3 or 4 students and retelling it to the rest of the class, taking turns within the group so that each member contributed 1 or 2 sentences before passing the floor. During each story retelling, the teacher recast any errors involving the two functions of articles being targeted.

Sheen found no differences in the recast and control groups’ improvement in grammatical accuracy when she included all participants in her analysis without considering anxiety. Interestingly, however, on the speeded dictation and narrative-writing post-tests, the learners performed differently according to their levels of anxiety, with the low-anxiety recast learners outperforming both the low-anxiety control learners and the high-anxiety recast learners. The low-anxiety participants who had received recasts also produced more modified output and repaired their errors at a higher rate than the high-anxiety recast participants did. In other words, effects of error correction may initially have been obscured by an ID variable: Recasts were apparently usable and beneficial—for learners with low anxiety. Sheen interprets these results as confirming DeKeyser’s (1993) findings.
In drawing conclusions based on Sheen’s (2008) findings, it is important to keep in mind that there were limitations in the design of the study and unknown factors that may have affected the results. For one thing, Sheen does not report how many high- versus low-anxiety learners there were in each intact class, nor does she report anything about the composition of the groups of students who worked together on the treatment tasks (e.g., whether of high or low anxiety or a mix). Moreover, although she reports that the recast and control groups contained similar proportions of learners from the various L1 backgrounds, she does not report how the learners’ anxiety levels may have differed according to L1; thus, it is not possible to know whether perhaps high- or low-anxiety learners tended to speak L1s which either contained articles or did not. Sheen does note that the learners in the low-anxiety recast group (who showed the most improvement) also happen to have received a greater number of recasts than those in the high-anxiety recast group (59 versus 48), and she points out that in future research it would be important to include a comparison group that completed the story-retelling activities without receiving feedback in order to establish more definitively that it was the recasts themselves that influenced learning and not simply the fact of having engaged in the treatment tasks.

Finally, it is not clear how to account for the fact that the low-anxiety recast group did not significantly outperform the high-anxiety control group, whose scores increased somehow from pre- to post-test as well. Sheen admits also that “there is no obvious explanation” for why the high-anxiety control group might have performed better than the low-anxiety control group (p. 866). In her review of the literature, however, she points to debates over whether anxiety is necessarily debilitating or facilitative of SLA in all circumstances. With this in mind, it is possible
to speculate that the effects of anxiety might have cut both ways in her study: Learners who were prone to being anxious about speaking in public and who were placed in precisely that sort of situation during the treatments with recasts might have been hampered by the intense anxiety they experienced, whereas learners prone to being anxious who were not placed in that situation (and instead took the pre- and post-tests without receiving recasts during public speaking activities) might independently have experienced some positively motivating effects of mild anxiety. Whatever the true explanation, and even though the research design may not have dealt adequately with potentially intervening variables, the results are at least suggestive of an ID-treatment interaction that may be worth exploring further in more rigorously controlled experiments.

Another classroom-based ATI study was conducted by Erlam (2005) with learners of L2 French in a New Zealand high school. She assigned 92 students to 3 experimental groups and a control, and measured their language analytic ability (MLAT IV, Words in Sentences), phonemic coding ability (PLAB Sound Discrimination), and phonological short-term memory (PSTM, using a self-designed, group-administered written test) to explore whether those cognitive capacities would be differentially related to performance under different conditions of L2 exposure. The linguistic target was direct object pronouns. One instructional method, referred to as deductive instruction, involved providing explicit rule explanations and form-focused production activities which “allowed [the learners] time to think and apply the rules they had learnt” (p. 154). Participants in this group were given a chart of all direct object pronoun forms, as well as corrections after completing the activities. There was also an inductive instruction method, which involved encouraging the students to engage in active hypothesis testing without
receiving any explicit rule explanations, followed by production and input-based consciousness-raising activities. These participants were not given the chart of direct object pronouns and were not told to search for patterns. Lastly, a structured-input instruction method involved providing explicit rule explanations and input-based consciousness-raising activities without production practice, “loosely based on descriptions of input processing” approaches (p. 155). Some of the tasks involved identifying errors in spoken and written input, and the learners were provided with correct answers and explanations. All three experimental groups received explicit negative feedback, and each engaged in three 45-minute sessions with the researcher. Their listening and reading comprehension and oral and written production were assessed in a pre-test/post-test/delayed-post-test design.

Examining the correlations between aptitude measures and testing gains (both immediate and delayed) for each group separately, Erlam found the following: In the deductive instruction group, which tended to outperform the other three (pp. 161, 163), there were no statistically significant correlations for language analytic ability or PSTM, but there were 2 strong and statistically significant correlations for phonemic coding ability (a positive one with the learners’ immediate listening comprehension and a negative one with their frequency of producing direct object pronouns in speech on the immediate post-test). In the inductive instruction group, there were no statistically significant correlations for phonemic coding ability or PSTM, but there were 3 strong and statistically significant correlations for language analytic ability (positive ones with immediate listening comprehension and delayed written production, and a negative one with frequency of delayed oral production). In the structured input group, there were no statistically significant correlations for phonemic coding ability, but there were 2
strong and statistically significant positive correlations each for language analytic ability and PSTM (in both cases with the learners’ immediate and delayed written production).

Erlam interprets these results as providing “little evidence” for a relationship between aptitude and L2 gains in the deductive group (p. 162). Bearing in mind that the learners in that group outperformed those in the other groups, she reasons that instruction with explicit rule explanation and production practice “tends to benefit all language learners” and “seems to minimize or level out any effect that individual differences in language aptitude may have with respect to instructional outcomes” (p. 163), supporting the notion that more scaffolding and information can sometimes serve as an equalizer (Skehan, 1998). Meanwhile, drawing attention to the relationships between language analytic ability and gains in the inductive instruction group, Erlam endorses the idea that analysis-oriented learners might be better able to form their own hypotheses and deal with less structured material (p. 165). Relationships between language analytic ability (and PSTM) and written production gains were also found, however, for the structured input group who had received explicit rule explanations. In that case, Erlam focuses on the fact that instruction in that condition did not involve production; that is, she argues that students with greater grammatical sensitivity and PSTM may have been better equipped to show gains in written production despite not having practiced it previously.

Given that the number of statistically significant correlations found in each condition was within the narrow range of 2 to 4, it may not be clearly justified to state, as Erlam does, that “deductive instruction that gives students opportunities to engage in language production minimizes any effect that individual differences in learner aptitude may have with respect to instructional outcomes” (pp. 167-8). Nonetheless, it is certainly interesting to reflect on the
differing patterns of correlations across groups: An aptitude auditory in nature was related to listening comprehension gains in a group that may not have practiced listening much; language analytic ability was related to performance on a variety of outcome measures in a group that had not been provided with rules; and written measures of PSTM and language analytic ability were related to written production gains in a group that had not practiced production. In the first case, perhaps this had to do with the nature of the outcome measure; in the second, with the nature of the learning processes; and in the third, with both.

A final classroom-based study, by Fujii (2005), lends further support to theoretical proposals regarding the multi-componential nature of aptitude, the existence of different aptitude profiles, and the idea that specific components of aptitude will be differentially related to specific dimensions of L2 performance. One of the major contributions of Fujii’s study, from her perspective, is that she was able to “move beyond more traditional correlational approaches” (p. 228) and “link SLA at the process level with aptitudinal constructs,” as recommended by Dörnyei and Skehan (2003, p. 599), through detailed qualitative analyses of learners’ stimulated recall protocols, learning journals, and modified output during interaction. The participants were 15 L1-Japanese learners of L2 English. Fujii measured their language analytic ability using the Language Aptitude Battery in Japanese (LABJ), their phonological short-term memory, and their working memory in L1 and L2, and she analyzed their L2 production with regard to its fluency, accuracy, syntactic complexity, lexical variety, and lexical sophistication. It was not an ATI study in the sense of comparing aptitude-learning relationships across different instructional conditions; rather, dividing the learners into high and low groups
for each aptitude component, Fujii was able to explore relationships among several dimensions of aptitude, the learners’ orientations to linguistic form, and various aspects of L2 production.

Fujii did not find any relationships between the aptitude components and the learners’ fluency, syntactic complexity, or lexical variety. However, she did find that lower language analytic ability was associated with greater orientation to form, higher language analytic ability was associated with an orientation only to task content and (perhaps correspondingly) lower morphosyntactic accuracy, higher PSTM was associated with greater syntactic accuracy, and higher WM was associated with greater lexical sophistication. Qualitative analyses revealed that the learners differed with respect to their L2 production priorities as well as with respect to their ability to prioritize multiple aspects of language production, in the latter case in relation to their WM capacities. Having discovered that orientation to form was often triggered by gaps in learners’ linguistic abilities, Fujii pointed out, optimistically, that low-aptitude learners’ problems communicating “opened doors to opportunities” for focus on form (p. iv). However, she also noted that, even when presented with such chances, learners’ strategic decisions and/or capacity limitations could make it so that they “may not necessarily capitalize on those opportunities in a way that enhances the accuracy and complexity of their speech” (p. 211).

In sum, as shown in the other classroom-based studies as well, the ways in which individual-difference variables interact with different instructional conditions and aspects of L2 performance can be exceedingly intricate, and it is crucial for researchers to be aware of the possibility of higher-order interactions which may obscure the sorts of relationships between learners’ abilities and features of the learning environment that might otherwise be expected.
Laboratory-based studies of aptitude-treatment interactions

An attraction of laboratory-based studies, particularly those involving (semi-)artificial grammars, is the amount of control they enable researchers to have over L2 exposure and other mediating variables. In one such study, comparing implicit and explicit instructional conditions, de Graaff (1997) measured specific aptitude components to investigate whether learners’ performance could be predicted to a greater extent in one condition than in the other. The participants were 54 Dutch-speaking undergraduates who engaged in 15 hours of computerized self-study activities (10 lessons of an hour and a half each, at a rate of 2 lessons per week) in eXperanto, a version of the language Esperanto modified by de Graaff to mimic certain features of Spanish. The linguistic targets, intended to represent simple and complex morphological and syntactic phenomena, were (1) noun pluralization involving vowel harmony (simple morphology), (2) the position of negative elements (simple syntax), (3) imperative formation involving formality and affirmativity distinctions (complex morphology), and (4) the position of unstressed versus stressed/topicalized pronominal objects (complex syntax).

During the lessons, the learners had to translate vocabulary, attempt to comprehend dialogues whose Dutch translations were then provided, make form-meaning connections in input-based activities, fill in blanks within sentences, and produce sentences based on clues given in Dutch. Both groups were given immediate feedback regarding the correctness of their responses and were provided with the right answers as necessary. Then, at the end of each session, all missed items were repeated. Learners in the explicit condition received explicit rule explanations both at the outset and as part of the feedback, whereas in the implicit condition, rules were never provided. The learners were tested following lesson 5, again after lesson 10,
and then once again 5 weeks later. Achievement measures included a fill-in-the-blank task, a grammaticality judgment test (GJT) with time pressure, and a GJT without time pressure in which the participants were additionally asked to make corrections. To assess their aptitude, de Graaff used Dutch versions of MLAT Part IV (grammatical sensitivity) and Part V (rote memory), as well as a test measuring the participants’ capacity to infer the meanings of eXperanto words in context. He then collapsed all of these measures into a single aptitude composite score.

Some of de Graaff’s main findings were that participants in the explicit condition outperformed those in the implicit condition on all 4 types of test across all 3 occasions of testing, and that aptitude tended to be statistically significantly related to test performance for learners in both conditions, implicit and explicit, on all test types except the GJT without time pressure. In other words, he found no evidence for aptitude-treatment interactions. De Graaff also reports that there was “no evidence [for] a differential effect of explicit instruction depending on the variables complexity and morphology/syntax” (p. 249); however, he did find that the difference between the implicit and explicit groups’ performance was greater on the complex syntactic structure than on the simple one (i.e., explicit instruction provided more of an advantage when the syntax was more complex). He suggests that the reason a similar effect of complexity was not found for the morphological targets may have been because the ‘simple’ noun pluralization rule, involving vowel harmony, proved to be more difficult for participants in the implicit condition than had been expected; thus, explicit instruction provided an advantage not only on the complex target, but also on the (not-so-simple) ‘simple’ target. Since de Graaff does not report the correlations between achievement measures and aptitude subtest scores, indicating only that “correlations... tended to be significant” for grammatical sensitivity and
lexical inferencing, but not for rote memory (p. 261), it is not clear whether perhaps specific components of aptitude may have been differentially related to performance across the implicit and explicit conditions.

This last issue, of course, has driven a good deal of Robinson’s research into aptitude complexes and ATI. A hypothesis in one of Robinson’s experimental studies of explicit versus implicit conditions was that, “[b]ecause different learning conditions may bias participants to rely more on memory… or on grammatical sensitivity…, different components of aptitude could trigger awareness [of language rules] under each condition” (1997b, p. 56). As in de Graaff’s study, Robinson (1997b) found that aptitude was related to performance in explicit and implicit conditions; however, this was not the case for both types of implicit condition he tested, and, as will be explained in greater detail below, the specific components of aptitude that showed associations with awareness and learning patterned differently across conditions.

In Robinson’s (1997b) experiment, 104 intermediate-level learners of L2 English were assigned to four different conditions: (1) implicit, in which they were told that they would be performing a memory activity and were asked questions about the locations of words in the input sentences; (2) incidental, in which they were told that they would be performing a comprehension activity and were asked yes/no questions about the meanings of the input sentences; (3) rule-search, in which they were asked to find the rules exemplified by the input sentences and then, during the treatment, were asked whether they were still looking for rules and whether they had found any; and (4) instructed, in which they were given written and oral explanations of the rules, were encouraged to apply them to the input sentences, and were asked rule-related yes/no questions about the sentences’ grammatical forms, with the
possibility of consulting the written explanations as needed. The two linguistic targets were (1) subject-verb inversion in sentences with fronted adverbials of movement or location (e.g., *Into the house ran John*, considered to be an ‘easy’ rule), and (2) the formation of pseudo-clefts of location (e.g., *Where John lives is in Chicago, not in New York*, considered a ‘hard’ rule).

During 2 computer-based training sessions, the participants were exposed to 40 sentences, 20 involving the easy rule and 20 the hard rule, and were allotted 10 seconds for each. Robinson measured learning with a 40-item GJT, and he measured aptitude using Part IV (grammatical sensitivity) and Part V (rote memory) of the MLAT (in English, despite the fact that this was the participants’ L2). In order to assess the participants’ awareness of the rules, Robinson administered a post-task questionnaire with 3 items, asking whether they had noticed any rules, had been looking for rules, and could describe the rules. Their answers to these questions were coded as simply affirmative or negative for the easy rule and the hard rule, “on the basis of any attempt at a metalinguistic description... or the presentation and comment on a relevant exemplar” (p. 63) without distinguishing between detailed versus partial or accurate versus inaccurate descriptions (p. 76).

Robinson found that none of the aptitude or awareness measures were related to learning in the incidental group and that simply noticing rules was not related to learning in any group. However, grammatical sensitivity was related to learning both of the rules in the instructed and implicit groups, and also to learning the easy rule in the rule-search group. Rote memory was related to learning both of the rules in the instructed group and to learning the hard rule in the rule-search group. Reports of having looked for rules were associated with learning in the implicit group, and the ability to verbalize rules was associated with learning
both of the rules in the implicit group and the hard rule in the rule-search group. These results are summarized in Table 2.1.

**Table 2.1.** Aptitude components and awareness reports associated with learning easy and hard rules across conditions in Robinson (1997b)

<table>
<thead>
<tr>
<th></th>
<th>Grammatical sensitivity</th>
<th>Rote memory</th>
<th>Noticing</th>
<th>Looking for rules</th>
<th>Ability to verbalize</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Implicit</strong></td>
<td>Easy, Hard</td>
<td>X</td>
<td>X</td>
<td>Easy, Hard</td>
<td>Easy, Hard</td>
</tr>
<tr>
<td><strong>Incidental</strong></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>Rule-search</strong></td>
<td>Easy</td>
<td>Hard</td>
<td>X</td>
<td>X</td>
<td>Hard</td>
</tr>
<tr>
<td><strong>Instructed</strong></td>
<td>Easy, Hard</td>
<td>Easy, Hard</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Robinson (1997b) interprets these aptitude-treatment interactions in light of claims by Krashen (1981, 1982, 1985) and Reber (1989, 1993; Reber et al., 1991) to the effect that IDs should not predict learning in incidental/implicit conditions and that aptitude/awareness should be associated with learning only when the rules are easy enough to grasp. Robinson acknowledges that the relationships between aptitude components and learning in the rule-search condition seem to support Reber’s perspective: Looking for rules can help if they are easy and learners have sufficient sensitivity to grammar, but when the rules are harder, it is memory ability that plays a role. However, other observed relationships lead him to question Krashen and Reber’s proposals. Robinson considers it telling, for example, that the correlations between grammatical sensitivity and accuracy were actually strongest in the implicit condition ($r=.69$ for the easy rule and $r=.75$ for the hard rule, compared to correlations between $r=.28-.60$.
in the other groups) and that, despite an emphasis on memory in the instructions for that condition, memory ability did not predict those learners’ performance. Moreover, learners with higher levels of grammatical sensitivity in the implicit condition were more likely to have looked for rules and to have been able to verbalize them, and those awareness reports were in turn positively associated with test performance.

Before following Robinson into any conclusions about the influence (or not) of IDs on the ‘implicit’ learning of ‘hard’ linguistic targets, it seems important to step back and reassess the appropriateness of these labels in the context of his study. Both linguistic targets in Robinson’s experiment (i.e., pseudo-clefts and subject-verb inversion with fronted adverbials) involved word order, and the questions posed during training in the so-called ‘implicit’ condition were designed to draw participants’ attention to the locations of words. This may have increased the explicitness of the treatment and reduced the practical complexity of the stimulus domain. Furthermore, even without such pointers, Robinson’s ‘hard’ rule seems less complex (impressionistically speaking) than the multiple rules derivable from one of Reber’s artificial grammars, whose possible ‘sentences’ include variations of the strings T[S]XS, T[S]XX[[T]VPX]VV, T[S]XX[[T]VPX]VPS, P[[T]VPX]VV, and P[[T]VPX]VPS (i.e., strings of surface forms without associations to meaning; see Figure 2.1 for an example).
Reber (1989) emphasizes that implicit learning (and research into it) requires “arbitrary stimulus domains with complex, idiosyncratic structures” (p. 219); “if the system in use is too simple, or if the code can be broken by conscious effort, then one will not see implicit processes” (p. 220). When a learning domain is complex enough, on the other hand, “explicit processing of complex materials has a decided disadvantage in relation to implicit processing” (p. 223), and learners who attempt conscious strategies without helpful external scaffolding will perform worse. In explaining this, Reber describes a study by Howard and Ballas (1980), in which explicit approaches were detrimental to performance “when there was no systematically interpretable pattern…. under conditions of semantic uninterpretability,” but facilitated performance when “the underlying factors that represent the grammar were rendered salient... [and a] semantic component focused the subjects on the relevant aspects of the patterned stimuli” (Reber, 1989, p. 223; see also Moeser & Bregman, 1972; Morgan & Newport, 1981). Reber also describes one of his own studies, in which participants performed poorly using an explicit approach when complex stimuli were presented haphazardly, but “the simple expedient of arranging the
exemplars of the grammar according to their underlying form produced... instructional facilitation” because “increasing the salience of the relationships between symbols increases the effectiveness of subjects’ attentional focus.... It alerts them to the kinds of structural relations that characterize the stimuli that follow and permits appropriate coding schemes to be implemented” (pp. 223-4). Relating this to the issue of aptitude-treatment interactions, Reber would fully agree that individual differences (e.g., in IQ) can play a role in explicit approaches to learning (see, e.g., Reber et al., 1991).

Robinson (1997b) acknowledges several points such as these in his review of the literature (pp. 51-2); however, when drawing conclusions, he seems perhaps not to be recognizing the potential differences in complexity between his ‘hard’ rule and the set of rules represented in Reber’s stimuli, and perhaps not to be recognizing certain potentially attention-drawing features of the ‘implicit’ condition—features which are crucial to consider if one wishes to theorize about the nature of aptitude-treatment interactions. 7 Robinson states that his participants’ reports of having searched for rules in the implicit and incidental conditions “[cast] doubt on the claim that learning under these conditions proceeds unconsciously” (p. 77), and he states that the relationships between awareness and learning in the implicit and incidental conditions “cast doubt on... claims about the superiority of unconscious learning processes in complex stimulus domains” (p. 80). However, Reber might not actually be inclined to claim that learning in Robinson’s ‘implicit’ condition should proceed unconsciously or that unconscious learning should be superior in that condition; it may not have been complex enough. In other

---

7 Though beyond the scope of the present discussion, there is potentially also the issue of learners’ goals, motivations, and inclinations to search for rules in real-language experiments like Robinson’s versus in artificial-grammar experiments with strings of meaningless letters like Reber’s.
words, it is not obvious that Robinson’s results cast doubt on Reber’s claims about implicit learning in complex domains where ‘implicitness’ and ‘complexity’ are operationalized as intended.

A point of the current discussion is to emphasize, as Robinson (2001, 2005a) himself does, that it is critical to analyze the characteristics of learning targets and to do so very carefully in relation to characteristics of treatments that might draw attention to relevant aspects of them. Post-hoc assessments of the actual complexity and processing requirements of targets and tasks may sometimes be necessary, and—importantly—the types of aptitude-treatment interactions that are observed can help researchers to make inferences about learners’ approaches. If SLA researchers wish to draw conclusions about how influential language analytic ability might be when L2 learning conditions are ‘implicit’ or ‘explicit’, for example, it may be helpful to do some post-hoc labeling of treatment conditions as well. It may turn out, after all, that participants in a treatment which was intended to promote the implicit learning of a complex target actually take a conceptually driven rule-search approach and find it easier than the researcher expected. In such a case, the researcher can certainly report what was intended in the experimental design, but may wish to avoid the term ‘implicit’ as a probable misnomer, to hedge about the linguistic target’s actual level of complexity, and to hold off on drawing conclusions about implicit learning of complex targets.

In reasoning about the differing aptitude-awareness-learning relationships across the other treatment conditions of his study, Robinson’s (1997b) analyses make sense. He points out that it is unsurprising that reports of having searched for rules were not associated with better performance for learners in the rule-search condition (who had been told to look for rules) or
for learners in the instructed condition (who had already been given the rules). As a matter of fact, in the instructed condition, learners who reported looking for rules performed worse than those who did not, and Robinson speculates that this may be because the ones who reported looking for rules did so because they could not remember them (p. 78). Reflecting on the characteristics of the incidental condition, and noting that neither grammatical sensitivity, nor memory ability, nor any type of awareness report was associated with learning, Robinson acknowledges (1997b, pp. 75-6) that the results appear to support Krashen’s (1981, etc.) claim that the acquisition which occurs while processing L2 input for meaning is unaffected by IDs in aptitude, as measured by the MLAT. Robinson (2002) explains, however, that another type of aptitude, more closely related to the specific L2 processing demands of the incidental condition, might have been associated with performance: “working memory—a fundamental capacity drawn on in ‘noticing’ grammatical information while primarily oriented to processing for meaning” (p. 229). He investigated precisely this in a later experiment.

In Robinson’s (2002) study, 55 Japanese-speaking undergraduates participated in a computer-based experiment devised to investigate explicit, implicit, and incidental learning in relation to individual differences in IQ, aptitude, and working memory. A control group of 17 participants completed only the tests and some initial vocabulary training, whereas the other 38 additionally participated in all 3 treatment conditions in a repeated-measures design. The explicit and implicit conditions, completed on the first day in about 40 minutes, were intended to be comparable to conditions of those types in Reber’s work. For the explicit condition,

---

8 Robinson (1997b) nonetheless maintains that since “large numbers of Incidental participants did claim to have noticed and looked for rules... conscious processes were clearly implicated in whatever incidental learning occurred” (p. 82).
Robinson employed a series-solution problem task (Reber et al., 1991) in which the participants were asked to choose which letter of the alphabet would most logically complete a patterned string of letters. In the implicit condition, participants also saw strings of (different) letters, this time generated by an artificial grammar (as in Figure 2.1). They were told that they would be engaging in a memory task and should write down each string after it was presented. Then, after training, they were informed that the strings had followed a set of rules and, presented with new strings, were asked to decide if they followed the same rules. The incidental condition, which involved learning 3 rules of Samoan from scratch, took place later over several days. First, the participants were trained on Samoan vocabulary; then, they completed 2 training sessions on separate days, followed by immediate, 1-week-delayed, and 6-month-delayed post-tests. Over the course of the training sessions, the participants were exposed to 450 Samoan sentences (150 exemplars of each rule type) in 10 trials of 15 minutes each, in each of which 45 sentences were shown for 10 seconds each. No grammatical explanations were provided; the instructions simply asked the participants to try to understand the meanings of the sentences and then to respond to yes/no comprehension questions.

The achievement tests of Samoan learning included an unspeeded computer-based GJT, an aural GJT using the same sentences, and a guided sentence production task, in which the participants were given words written on slips of paper and had to arrange them to form sentences. The participants also completed an awareness questionnaire similar to that used in Robinson’s (1997b) study. To measure individual differences in IQ, aptitude, and WM, Robinson employed a Japanese version of the short form of the revised Wechsler Adult Intelligence Scale (WAIS-R), the LABJ, and a reading-span task (Osaka & Osaka, 1992), respectively. The linguistic
targets were (1) an ergative marker, \( e \), which precedes the subject of a transitive verb; (2) a locative marker, \( i \), which precedes a specified location; and (3) noun incorporation, in which Samoan’s predominantly VSO word order is changed by incorporating the object into the verb.

Considering first the relationships between IDs and learning in the explicit and implicit conditions, Robinson (2002) found no associations with WM, but he did find that IQ and aptitude were positively associated with performance on the explicit task, whereas IQ was negatively associated with performance on the implicit task—possibly because the participants “adopted an explicit code-breaking set…. and since the rules governing the implicit learning task… are extremely complex, this had negative effects” (pp. 253-4). (In other words, analogously to above, higher IQ was associated with less successful attempts at explicit learning in a so-called ‘implicit’ condition where stimuli were too complex for explicit learning to work.)

In analyzing the results for the incidental condition, Robinson found that, overall, there was no association between the learners’ performance on the immediate post-tests and their IQ, aptitude, or awareness, nor were there relationships with their previous performance in the implicit or explicit conditions. However, Robinson did find statistically significant positive correlations between WM and the learners’ overall performance on the immediate and 1-week delayed aural GJT's and the 1-week and 6-month delayed production tests. He also found statistically significant positive relationships between the learners’ overall performance on the 6-month delayed production test and their aptitude and reports of awareness on the second questionnaire he administered. In light of these findings, he concludes that learning was “most sensitive to the ID measures that most closely matched the abilities hypothesized to be drawn on during incidental learning task performance, i.e., aptitude and working memory” (p. 256).
To gain greater insight into the nature of these relationships, Robinson (2002) examined learners’ performance on each of the different rule types separately. In doing so, he found that they tended to enjoy greater success with the structures that, he reasoned, seemed more likely to involve attending to form and meaning: The participants displayed “clear evidence of generalizable knowledge” of the locative structure on the GJTs and performed well with noun incorporation on the production tests despite not showing knowledge of its restrictions on the GJTs. However, there was “no evidence, beyond memory for old instances, of learning the ergative rule” (p. 249), which involved “semantically opaque and communicatively redundant” morphology (p. 257). Looking at the relationships between test performance for each target and the various ID measures, Robinson found the following (also displayed in Table 2.2):

- Performance with the **locative** was statistically significantly related to **aptitude** on the 6-month-delayed computerized GJT, as well as on the immediate and 6-month-delayed production tests. However, no statistically significant relationships were found for the aural GJT, or for any of the other ID measures (IQ, WM, or awareness).

- Performance with **noun incorporation** was statistically significantly related to **aptitude** and **IQ** on the immediate aural GJT, to **WM** on the immediate and 1-week-delayed aural GJT, and to **awareness** on the 1-week-delayed computerized GJT. No statistically significant relationships were found for the production test.

- Performance with the **ergative** structure was statistically significantly related to **aptitude** on the 6-month-delayed production test, to **IQ** on the immediate production test, to **WM** on the 1-week-delayed aural GJT and 1-week-delayed
production test, and to awareness on the 6-month-delayed production test. No statistically significant relationships were found for the computerized GJT.

Table 2.2. Tests showing statistically significant correlations between IDs and L2 performance for each rule type in Robinson (2002)

<table>
<thead>
<tr>
<th></th>
<th>Aptitude</th>
<th>IQ</th>
<th>WM</th>
<th>Awareness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Locative</strong></td>
<td>Written GJT T3</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>(.47)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Production T1, T3</td>
<td>(.47, .50)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Noun incorporation</strong></td>
<td>Aural GJT T1</td>
<td>Aural GJT T1</td>
<td>Aural GJT T1, T2</td>
<td>Written GJT T2</td>
</tr>
<tr>
<td></td>
<td>(.54)</td>
<td>(.46)</td>
<td>(.41, .52)</td>
<td>(.35)</td>
</tr>
<tr>
<td><strong>Ergative</strong></td>
<td>Production T3</td>
<td>Production T1</td>
<td>Production T2, T3</td>
<td>Production T3</td>
</tr>
<tr>
<td></td>
<td>(.44)</td>
<td>(.55)</td>
<td>(.42)</td>
<td>(.43)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Aural GJT T2, T3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.49)</td>
<td></td>
</tr>
</tbody>
</table>

Key: T1 = Immediate post-test, T2 = 1-week delayed post-test, T3 = 6-month delayed post-test
The numbers in parentheses are correlation coefficients (r).

Somewhat inexplicably in relation to these results, Robinson states that “the effects of WM on incidental learning are most clearly apparent (not surprisingly) on the rules which can be learned” (p. 256). However, considering that the participants displayed “clear evidence of generalizable knowledge” for the locatives (where no relationships with WM were found) and “no evidence... of learning the ergative rule” (where relationships with WM were found) (p. 249), this statement does not seem to be accurate.

Perhaps one of the main lessons to take from this (and other ATI research) is that the more interactions there are in a study, the more difficult it can be to tease them apart, keep
them straight, and interpret them. However, it is also interactions that will increase the
precision and validity of our theories of SLA. In discussing the results of his research as a whole,
Robinson (2002) makes an excellent point in stating:

IDs in cognitive abilities interact with rule complexity and rule type during
incidental learning. The true nature of the influence of IDs on incidental learning
is therefore to be found in the specific details of such interactions... and this is
not, ultimately, reducible to general statements that IDs do, or do not, affect
global incidental, explicit, or implicit learning (p. 260).

Applying this proposition to Robinson’s interpretations of his empirical results, some tempering
of other statements he makes may be in order. For instance, Robinson concludes that “working
memory appears to have the strongest influence on incidental Samoan learning” (p. 241),
whereas “aptitude and awareness are only weakly related to incidental learning” (p. 243).
“[O]verall,” he states, “the LABJ measure of aptitude is a weak predictor of successful incidental
learning... [and may] need to be revised... to capture the cognitive abilities drawn on” in that
situation (p. 261). These ideas may have seemed reasonable when examining the learners’
performance on all 3 linguistic structures combined; after all, as we saw above, a relationship
between aptitude and overall test scores was found only on the 6-month-delayed production
test, whereas relationships between WM and overall test scores were found on 4 achievement
measures. However, when examining the learners’ test performance for each linguistic target
separately, as shown in Table 2.2, there were 5 statistically significant correlations with
aptitude and 4 with WM. That is, perhaps the LABJ is not such a weak predictor of incidental
learning after all; analyses (and interpretations) simply have to be fine-grained enough.
Situations such as these seem to be par for the course in L2 aptitude research. In study after study (e.g., DeKeyser, 1993; Sheen, 2008; one could speculate possibly even de Graaff, 1997), what initially looks like no relationship turns out to be composed of several smaller and more intricate ones. In view of this, it seems sensible to argue that if researchers have data that can speak to more narrowly defined relationships, these should be examined closely before drawing any conclusions about overall relationships or an apparent lack thereof. In some cases, as in Lado’s (2008) study to be discussed below, it may be that not many aptitude-treatment interactions are found; nonetheless, it is important to check.

Like Robinson (2002), Lado (2008) used multiple outcome measures to explore potentially differing relationships among various aspects of aptitude and learning across different treatment conditions. In this experiment, described earlier in the review of research on computer-mediated feedback, an advantage was found for explicit feedback over implicit feedback in helping ab initio learners of Latin as an L3 learn to use morphological cues for assigning semantic roles in sentential contexts. The outcome measures included tests of written and aural interpretation, grammaticality judgments, and sentence production, and the aptitude measures included the MLAT and tests of WM and phonological short-term memory (PSTM) in the learners’ L1 (English) and L2 (Spanish). In brief, according to Lado, the “cognitive capacity measures predicted language learning success to different degrees across [outcome] measures” (p. 249), but “the Treatment effect was only marginally affected by aptitude and WM capacity” (p. 225).

In order to investigate whether there had been any aptitude-treatment interactions, Lado’s approach was first to perform repeated-measures analyses of variance (ANOVAs) testing
for treatment effects over time (i.e., time-by-treatment interactions) irrespective of IDs, and then to perform separate repeated-measures analyses of covariance (ANCOVAs) for each possible combination of outcome and aptitude measures, using the achievement test scores as the dependent variable and entering MLAT, WM, L1 PSTM, and L2 PSTM scores (one at a time) as covariates. Whenever there was a difference in the results of the ANOVA and ANCOVA (e.g., if a previously statistically significant time-by-treatment interaction disappeared when holding ID scores constant), she investigated further by (1) running a one-way ANOVA comparing the treatment groups on that ID measure to check whether one of the groups might have had a cognitive advantage relative to the other; and (2) running correlations between ID scores and test performance for each group separately to evaluate whether the relationships between IDs and achievement differed across treatment conditions.9

Holding WM scores constant, Lado did not find differences between her previous ANOVA results and the ANCOVA results. Accordingly, she reports that working memory “did not seem to play a role in the effect that Type of Feedback had on language learning” (p. 215).

Controlling for MLAT scores, Lado was also led to infer that, for the most part, aptitude probably played a similar role across treatment conditions. However, she did identify one part of one outcome measure where the results of the ANOVA and ANCOVA differed: On the

9 It seems another approach might have been taken without using differences between the ANOVA and ANCOVA results as prerequisites for further analyses. Thomas et al. (2009) point out that, particularly in cases where a covariate explains more variance than a task effect (here, performance on the pre- vs. post-tests), a RM ANCOVA might weaken the task effect compared to when it is assessed using a simple RM ANOVA—“as if the between-subject covariate is taken to explain some of the variability in the (t1-t2) difference score” even though conceptually it should not. Their solution is to use a two-phase analysis in which they first use a RM ANOVA to evaluate any repeated-measures effects and interactions, and then use a separate RM ANCOVA to evaluate the main effect of the ID and any interactions with the ID. Possibly, instead of pursuing further analyses of IDs only if the ANOVA and ANCOVA results differed, Lado might have focused on Time*Treatment*ID interaction effects within RM ANCOVAs. Of course, the results she actually reports can still be reflected upon, but perhaps with the caveat, frequently admitted in any case, that they may not tell the whole story of the ATIs that existed.
participants’ written production of new items, when MLAT scores were included as a covariate, Lado found that a previously non-statistical time-by-treatment interaction ($p=.072$) reached statistical significance ($p=.048$), with the explicit feedback group showing greater improvement than the implicit group over time. Correlations between the participants’ MLAT scores and their written production of new items, performed separately for the explicit versus implicit treatment groups, were strong, positive, and statistically significant in both conditions (explicit: pre .26, post .62, delay .51; implicit: pre .39, post .48, delay .39). From all of this, Lado surmises that “participants with higher aptitude in the [implicit condition] were able to compensate for the lack of rules and perform at levels closer to the [explicit] group” (p. 260). In other words, she discovered more evidence for the benefits of explicit feedback as well as some apparent evidence for the helpfulness of L2 aptitude when learners are provided with less explicit information about a learning target.

Turning to relationships with the participants’ phonological short-term memory capacity, when Lado entered L1 PSTM as a covariate, she found a difference between the ANOVA and ANCOVA results for new items on the aural interpretation test which seemed to suggest that L1 PSTM may have contributed to the positive effects of explicit feedback. Looking at the correlations between L1 PSTM and aural interpretation performance for the implicit and explicit conditions separately, however, she found no statistically significant relationships for either group. Therefore, Lado reports that L1 PSTM did not appear to play a role in the different effects of implicit versus explicit feedback (p. 219). When controlling for L2 PSTM, Lado also found a very slight difference between the ANOVA and ANCOVA results for new items on the aural interpretation test which happened to straddle the alpha level ($p=.049$ for ANOVA versus...
Looking at the correlations between L2 PSTM and aural interpretation performance for the implicit and explicit conditions separately, she found that L2 PSTM was statistically significantly related to aural interpretation post-test performance on new items in the implicit group ($r = .28^*$), but not as strongly in the explicit group ($r = .19$). However, noting that the participants’ L2 PSTM capacities were also related to their L2 Spanish proficiency, she decided to explore this result further by running the correlations again with the participants’ levels of bilingualism held constant. In the end, Lado attributes the finding to the influence of L2 proficiency as opposed to L2 PSTM per se.

All in all, Lado concludes that her study did not produce strong evidence that cognitive capacities play different roles according to the type of feedback received (p. 262); under the statistical approach she took, she did not uncover clear aptitude-treatment interactions. However, importantly, as was also found to be the case in Fujii’s (2005) research, learners’ cognitive capacities were “differentially related to specific aspects of language performance” according to different outcome measures (p. 252). Although all of the ID measures predicted L3-Latin performance on at least some of the achievement tests in Lado’s experiment, the fact that many such relationships were found on sentence production, versus only a few on written interpretation, prompted Lado to suggest that more skills seem to be involved “under more demanding conditions” (p. 253). Participants’ MLAT scores, for instance, were related to their written production of old and new items on both post-tests, and also to their grammaticality judgments of old items on the immediate post-test, but not to their written or aural interpretation performance on any test. For working memory as well, associations were somewhat more consistently found between participants’ WM scores and their performance on
the GJT and production tests than they were between WM scores and written or aural interpretation. Table 2.3 summarizes the relationships that were found between IDs and performance for each outcome measure.

Table 2.3. Tests showing statistically significant correlations between cognitive capacities and L3 performance in Lado (2008)

<table>
<thead>
<tr>
<th></th>
<th>MLAT</th>
<th>WM</th>
<th>L1 PSTM</th>
<th>L2 PSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Written interpretation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Post</td>
<td>X</td>
<td>New</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Delay</td>
<td>X</td>
<td>Old</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>Aural interpretation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Post</td>
<td>X</td>
<td>X</td>
<td>Old, New</td>
<td>Old, New</td>
</tr>
<tr>
<td>Delay</td>
<td>X</td>
<td>Old, New</td>
<td>X</td>
<td>Old</td>
</tr>
<tr>
<td><strong>Grammaticality judgment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Post</td>
<td>Old, New</td>
<td>Old, New</td>
<td>X</td>
<td>Old, New</td>
</tr>
<tr>
<td>Delay</td>
<td>X</td>
<td>Old, New</td>
<td>X</td>
<td>New</td>
</tr>
<tr>
<td><strong>Production</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>Old, New</td>
<td>Old, New</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Post</td>
<td>Old, New</td>
<td>Old</td>
<td>Old</td>
<td>Old, New</td>
</tr>
<tr>
<td>Delay</td>
<td>Old, New</td>
<td>Old, New</td>
<td>X</td>
<td>Old, New</td>
</tr>
</tbody>
</table>
The issue of outcome measures may soon be attracting even more attention in studies of ATI. An in-press study by Révész examined whether 90 EFL learners’ performance on various types of assessment, following an instructional treatment with or without feedback on past-progressive forms, was mediated by their WM capacities. Révész discovered, among other things, that there were relationships between WM and L2 development in the feedback group, but not in the no-feedback group, and that within the feedback group, participants with higher reading spans were likely to show greater gains on written tests, whereas those with higher PSTM were likely to show greater gains on oral tests.

Révész hypothesizes that the latter set of results might be explained in at least two ways: with reference to learning processes (i.e., more than one route to L2 development) and/or with reference to the nature of the outcome measures (i.e., an aptitude/test-type interaction). In short, reasoning about the results in terms of learning processes, she considers it possible that (1) learners with high PSTM may have been able to maintain feedback in memory for a longer duration, enabling data-driven learning and proceduralization, which might have been accessible during the oral tests; and (2) learners with high reading spans (i.e., with greater ability to allocate attention to simultaneous storage and processing) may have been able to focus consciously on the feedback and develop declarative knowledge, which might have been accessible during the written tests. Alternatively, reasoning about the results in terms of the nature of the outcome measures, Révész suggests that (1) learners with high PSTM may have been able to retrieve recently processed chunks from short-term memory during the oral tests, as those tests repeatedly elicited the use of the same construction; and (2) learners with high
reading spans (measured by a test in the written modality) may have been better equipped to perform well on other written measures which likewise drew on literacy-related skills.

Where the results are clear-cut enough for this sort of detailed reasoning to be possible, this seems to be precisely what will help SLA researchers to draw theoretically interesting and practically useful conclusions from the aptitude-treatment interactions they find. Whether Révész’s findings are best explained with reference to learning processes or outcome measures or both, they substantiate Skehan’s (1998, 2002) and Robinson’s (2001, 2005a) emphases on the importance of taking fine-grained measures of a variety of cognitive abilities that may contribute differentially to L2 acquisition and performance under different conditions. Moreover, such findings might give rise to optimism regarding the idea (from Pellegrino and Glaser, 1979) that “the significant use of measures of intelligence and aptitude is not primarily for the purposes of prediction, but for indicating how intellectual performance can be improved” (p. 61, cited by Robinson, p. 378). The more L2 researchers and practitioners can discover about how IDs shape learning processes and performance under different task conditions, the better instruction can be adapted to learners’ characteristics.

We have seen from classroom-based studies of ATI that the multitude and complexity of variables naturally involved—including, for example, classroom anxiety (DeKeyser, 1993; Sheen, 2008) and learners’ priorities (Fujii, 2005)—can create higher-order relationships that obscure the more basic associations we might otherwise expect to find. We have seen from laboratory-based studies of ATI that increasing the complexity of our research designs (e.g., in terms of the number of treatment conditions, cognitive measures, outcome measures, and so on) can sometimes be disorienting. Researchers in ATI studies may indeed feel as though they are
entering Cronbach’s (1975) “hall of mirrors that extends to infinity” (p. 119). Decades ago, Snow
(1989) concluded that aptitude-treatment interactions were difficult to demonstrate clearly
because of their intricate nature and, correspondingly, not understood adequately to guide the
customization of teaching methods. However, as more is learned about SLA processes, and as
SLA researchers become more skilled at selecting ID measures, linguistic targets, task features,
and assessment instruments for principled reasons within well-defined frameworks, our results
should become easier to interpret, and our explanations should become more valid. Ideally, this
could become a virtuous circle in which, to a greater and greater extent, studies of “three-way
interactions between aptitudes, treatments, and psycholinguistic features of the learning
targets... provide much more insight into all three of these factors than the study of any one of
them in isolation can hope to accomplish” (DeKeyser, 2003, p. 337). It may also represent
another way in which researchers from different SLA subfields, such as generativists and
cognitive-interactionists, can contribute their expertise in complementary areas (e.g.,
knowledge of theoretical linguistics on the one hand, and knowledge of feedback and individual
differences on the other) and work toward a more multifaceted, intricate, and comprehensive
understanding of how languages can be learned most efficiently.
CHAPTER 3

REFLEXIVES IN A SECOND LANGUAGE

Theories of reflexives employed in empirical studies of their acquisition

For decades, learners’ interpretations of sentences containing reflexives (e.g., English himself, Japanese zibun ‘self’) were considered a suitable “microcosm” for testing Universal Grammar (UG)-based accounts of language acquisition (Thomas, 1991, p. 216). Well-defined (albeit sometimes inaccurate) generalizations about them guided numerous investigations into whether adult learners’ L2 grammars followed universal constraints (e.g., sensitivity to structure, such as c-command) and whether it was justified to argue for ‘direct access’ to UG (e.g., in cases where properties of the L2 seemed to be neither straightforwardly available in the input nor accessible via instantiation in the L1). Linguistic theories are constantly changing, however, even within particular approaches to SLA, and as White (1995a) points out, a “major problem” in SLA research has been that “the tests researchers devise to test one theory may fail to include material which is crucial to test the new or revised theory” (p. 65). This has certainly been the case historically for theories of reflexive binding, and the current state of affairs may be even more complicated: Developments in linguistic theory related to the goals and guiding principles of Chomsky’s Minimalist Program (e.g., 1995; see also Hornstein, Nunes, & Grohmann, 2005) have so profoundly challenged fundamental assumptions about binding that SLA researchers have (perhaps as a result) become rather quiet on what was once a central subject of linguistic inquiry.
The present study does not subscribe to a particular UG-based theory of the SLA of reflexives. However, the selection of target sentences was informed by a close examination of the structures used in previous empirical work on L2 reflexives, which has by and large been conducted by generative researchers. Since understanding the results of those studies requires at least a basic familiarity with their theoretical assumptions, three major UG-oriented approaches to explaining the properties and acquisition of reflexives will be reviewed here: Manzini and Wexler’s (1987, Wexler & Manzini, 1987) highly influential parameterized approach, Progovac’s (1992, 1993) Relativized SUBJECT approach, and a variety of related approaches involving movement at Logical Form (LF) (e.g., Cole, Hermon, & Sung, 1990; Cole & Sung, 1994; Pica, 1987).

In reviewing these approaches, what will be more important for the present investigation than a comparative evaluation of their technical details is an awareness of the counter-examples to generalizations regarding the behavior of reflexives that the theories have taken as foundational. As will become clear in the review of empirical research in the section following this one, some SLA researchers continue, even now, to interpret their results in light of these theories and to formulate new hypotheses about acquisition processes (e.g., involving universal morphological triggers) which would seem to rely crucially on the non-existence of such counter-examples. Perhaps the counter-examples will eventually be explained in ways that salvage the researchers’ assumptions, but in the meantime some skepticism seems warranted.
Manzini and Wexler’s parameterized approach

Manzini and Wexler (henceforth \{M,W\}) designed their theoretical proposals about reflexives with an eye to solving the ‘logical problem’ of language acquisition—the idea that, despite ‘impoverished’ input and a lack of negative evidence, children acquiring an L1 demonstrate knowledge that seems impossible to account for on the basis of imitation, analogy, etc. (Baker & McCarthy, 1981; Chomsky, 1986; Hornstein & Lightfoot, 1981; as explained by Thomas, 1991, p. 211). Thomas (1995) suggests that most early empirical studies (e.g., Broselow & Finer, 1991; Finer, 1991; Finer & Broselow, 1986; Hirakawa, 1990; Matsumura, 1994; Thomas, 1989, 1991, 1993; Yuan, 1994) were based on \{M,W\}’s proposals precisely because “explicit, testable hypotheses about L1 and L2 acquisition [could] be extrapolated from [them]” (p. 207). By the mid-1980s, the classical binding theory had become “largely untenable due to difficulties in accommodating evidence resulting from crosslinguistic research into binding” (Hicks, 2009, pp. 5-6). Accordingly, amending Chomsky’s Condition A (which stated, essentially, that an anaphor had to be bound in its governing category by a proper antecedent), \{M,W\} argued that, while the requirement of being co-indexed with a c-commanding NP was universal, the definitions of ‘governing category’ and ‘proper antecedent’ were parameterized and differed across languages. UG was understood as defining the limits of possible variation, and it was hypothesized that learners followed a Subset Principle whereby they started out assuming the most restricted values of these parameters and expanded their possibilities for anaphoric reference only on the basis of positive evidence.
{M,W}’s Governing Category Parameter (GCP) had 5 possible values determining the size of a reflexive’s local domain, and {M,W} assumed that it could be set differently for different lexical items within a language. Stated formally,

\( \gamma \) is a governing category for \( \alpha \) if \( \gamma \) is the minimal category that contains \( \alpha \) and

- a. has a subject, or
- b. has an INFL, or
- c. has a TNS, or
- d. has an indicative TNS, or

INFL, as explained by Thomas (1991, p. 217), was an inflectional element not present in nominal phrases or small clauses. Root TNS referred to the tense of a matrix clause, and the properties of an indicative TNS, as compared to a subjunctive, for example, were considered to be ‘inherently defined’. Reflexives in English (Type a) were considered to represent the most restricted setting of the GCP, whereby possible antecedents had to lie within the minimal category containing the reflexive and a subject and could not be located beyond an INFL. Japanese zibun (Type e) was considered to represent the least restricted setting, allowing antecedents at any distance within the sentence. Italian sé (Type b, INFL), Russian sebja (Type c, TNS), and Icelandic sig (Type d, indicative TNS) were taken as falling between these two extremes (Thomas, 1995, p. 207), and the scale was supposed to be one of hierarchical entailment; that is, a Type-e reflexive would allow all of the antecedent types allowed by Types a through d; a Type-d reflexive would allow all of the antecedent types allowed by Types a through c, and so on. Applied to language acquisition, if learners followed a Subset Principle,
this meant that children learning English, for example, would never allow antecedents beyond the most restricted Type-a setting because they would never encounter positive evidence suggesting that this was possible.

{M,W}’s Proper Antecedent Parameter (PAP) had only two possible values, stated formally as follows:

A proper antecedent for $\alpha$ is

a. a subject $\beta$, or


Under the more restricted setting (Type a, as with Japanese *zibun*), only syntactic subjects could serve as antecedents, whereas under the less restricted setting (Type b, as with English *himself*), antecedent NPs of any grammatical role were possible. According to the Subset Principle, learners were hypothesized to start out assuming a subject condition on antecedents. In Japanese, there would be no reason to expand this, whereas in English, positive input would indicate that other types of antecedent were allowed.

For a variety of reasons, researchers no longer assume that governing categories are parameterized, the Subset Principle is not believed to operate in SLA (see, e.g., Akiyama, 2002), and studies are no longer conducted using {M,W}’s framework. The most important shortcoming of that approach, according to Beck (1998), was considered to be that it “ignore[d] one of the most central of the crosslinguistic correspondences... namely, that monomorphemic reflexives permit LD [long-distance] binding whereas polymorphemic reflexives permit only local binding” (p. 9). Researchers also considered it a drawback that {M,W}’s theory did not
explain an apparent relationship between long-distance binding and subject orientation. As Thomas (1995) explains, \{M,W\}' recognized that certain settings of their GCP and PAP seemed to co-occur, but this was treated as descriptive as opposed to explanatory; the parameters were considered to be independent of each other (Wakabayashi, 1996, p. 270). Thomas (1995) describes some other prominent objections to \{M,W\}'s proposals as involving empirical and theoretical challenges to the Subset Principle (e.g., Kapur, Lust, Harbert, & Martohardjono, 1993; MacLaughlin, 1992; Saleemi, 1992); reactions against the arbitrarily stipulated nature of the features defining governing categories and proper antecedents (Reuland & Koster, 1991); and independent evolutions in binding theory whose central ideas were incorporated into counter-proposals to \{M,W\}'s approach.

Theories in the 1990s tried to account for observed links among reflexives' morphological characteristics, binding domains, and subject orientation. A similarity among these later approaches, as distilled by Wakabayashi (1996), was that they explained the subject-oriented long-distance binding properties of morphologically simple reflexives as being realized at Logical Form (LF), either “by movement (head-to-head—Pica, 1987; Cole et al., 1990; Cole & Sung, 1994—or adjunction—Katada, 1991) or transparency of recoverable functional categories (Progovac & Franks, 1992; Progovac, 1992, 1993)” (p. 292).

**Progovac’s Relativized SUBJECT approach**

During the 1990s, Progovac’s (1992, 1993) Relativized SUBJECT approach was adopted as a framework for several empirical studies of the SLA of reflexives (e.g., Bennett, 1994;
MacLaughlin, 1998; Wells, 1998). Assuming, in accordance with X-bar theory’s requirement for structural compatibility, that a reflexive and its antecedent had to have the same X-bar status, Progovac considered morphologically complex reflexives (such as *himself*) to be phrasal (XPs), consisting of a pronominal morpheme indicating person, gender, and number in the Specifier position, plus a morpheme meaning ‘self’ as the head of the phrase (e.g., \([NP_{\text{SPEC}} \text{him}] [N_{\text{self}}]\)). Morphologically simple reflexives (such as *zibun*), on the other hand, were heads (X0), consisting of only a single morpheme meaning ‘self’ (e.g., \([NP [N_{\text{zibun}}]\])). Building on Chomsky’s (1981) definition of a reflexive’s local domain as containing an accessible subject/SUBJECT (i.e., either an NP in Specifier position or a finite AGR [agreement], obligatorily co-indexed with the subject it governed [Haegeman, 2001, p. 222; Hicks, 2009, p. 24]), Progovac proposed that a Relativized SUBJECT for an XP reflexive could be the subject of a clause \([NP, IP]\) or the subject of an NP \([NP, NP]\), whereas an X0 reflexive could be bound to an AGR (also a head) which was co-indexed with a subject through Spec-head agreement. Aikawa (2002) explains that AGR would provide the \(\phi\)-features (person, number, gender) that a simple reflexive such as *zibun* needed in order to be interpreted (Bouchard, 1984).

Under Progovac’s approach, the binding of a polymorphemic XP reflexive was necessarily local because it could always find a potential antecedent within its clause. Potential antecedents for monomorphemic X0 reflexives, however, were also determined by the type of verbal agreement found in the language or in the relevant clause. Binding domains could differ depending on whether AGR in the language was ‘morphological’ or ‘anaphoric’ (Borer, 1989) and, in a language with morphological (overt) AGR, binding domains could differ also depending on whether the clause was finite or nonfinite. In cases where AGR was not realized
morphologically, long-distance binding across clauses was possible. In Russian, for example, which has morphological AGR in the language as a whole, the binding domain for the morphologically simple reflexive sebja was argued to extend outside only nonfinite clauses because infinitives lack morphological AGR. In Japanese, which has anaphoric AGR, the binding domain for zibun was argued to extend outside finite clauses as well as nonfinite ones because AGR is never morphologically realized. The idea here was that anaphoric AGRs could be linked to the AGRs in higher clauses through AGR-chains. Thus, zibun could be bound long-distance across finite clauses by being bound to a local AGR, which was in turn co-indexed with a higher AGR, which was co-indexed with a subject. In principle, the length of an AGR-chain, and therefore the size of the corresponding binding domain, was unrestricted. Non-subject NPs were not available as antecedents for zibun because they were not located in the Spec of AGR and therefore could not be co-indexed with AGR. In sum, for Progovac (1992, 1993), binding distance was related to the morphological characteristics of both reflexives and verbs, and the subject orientation of long-distance reflexives was related to their co-indexation with AGR.

Some potential problems with this approach (and, importantly, with its use by SLA researchers to support the idea of morphological triggers of binding behavior) include the fact that some monomorphemic reflexives (e.g., Serbo-Croatian sebe), as well as possibly some polymorphemic ones (e.g., Finnish hän itse), allow long-distance subjects, local subjects, and local objects as antecedents (Thomas, 1995, p. 217: attributed to Maling, 1986, p. 54; Bennett, 1993, pp. 30-31; and Reuland & Koster, 1991, p. 13, respectively; see also White et al., 1997, p. 165, note 9); the fact that Thai monomorphemic reflexive tuaʔeenŋ must be bound locally, while the polymorphemic reflexives tuachánʔeenŋ (first-person ‘self’), tuakháoaʔeenŋ (third-person...
'self'), etc., can take long-distance antecedents (Osatananda, 1993; Wilawan, 1991; as reported in Thomas, 1995, p. 211, note 4); the fact that some monomorphemic reflexives (e.g., Danish and Icelandic sig and Norwegian seg) have been reported actually to disallow local antecedents, obligatorily taking long-distance antecedents (Vikner, 1985, cited in MacLaughlin, 1998, p. 219, note 3; Sigurjónsdóttir & Hyams, 1992, cited in Akiyama, 2002, p. 40, and Hamilton, 1998, p. 296); the fact that some reflexives (e.g., Icelandic hann sjálfur, Norwegian ham selv) take only object antecedents (Hamilton, 1998, p. 296, citing a personal communication with Veturlidi Oskarsson, 1996); the fact that some long-distance reflexives (e.g., Dutch zich) do have ϕ-features in the sense that they are restricted to third-person referents (Reuland & Koster, 1991; cited in MacLaughlin, 1998, p. 198); and the fact that some reflexives (e.g., Japanese zibun-zisin and Norwegian seg selv) do not have ϕ-features despite being polymorphemic and, even so, can be bound only locally and only to subjects (Katada, 1991, p. 289; MacLaughlin, 1998, p. 220, note 5; Hellan, 1991, cited in Hamilton, 1998, p. 298, note 4). These examples are potentially problematic for the set of theoretical proposals to be discussed next as well since they likewise attempt to relate reflexives’ binding behavior to their morphological complexity.

**LF-movement approaches**

Like Progovac (1992, 1993), several other researchers have analyzed monomorphemic reflexives as heads and polymorphemic reflexives as NPs with a pronominal element in the Specifier position (e.g., Battistella, 1989; Cole et al., 1990; Cole & Sung, 1994; Huang & Tang, 1989, 1991; Katada, 1991; Pica, 1987). Though varying in their details, what these analyses have

---

in common with each other, but not with Progovac’s approach, is that they involved movement of reflexives at Logical Form (LF). As referentially dependent expressions, both polymorphemic and monomorphemic reflexives were argued to require movement at LF in order to be interpreted (or, as L. Jiang, 2009, puts it, as ‘defective’ elements, they had to move at LF in order to be licensed [p. 474]). Empirical studies of L2 reflexives adopting LF-movement approaches as their frameworks include Christie (1992), Thomas (1995), and L. Jiang (2009).

Thomas (1995) identifies Pica’s (1987) work as a starting point for the other analyses, all with the goal of “predict[ing] without additional mechanism the correlation between governing categories and proper antecedents” (p. 211). To account for the long-distance binding behavior of Chinese ziji, Battistella (1989) introduced the idea of iterative movement. The difference between the reflexive types was thought to be in their landing sites: Monomorphemic X⁰ reflexives (e.g., zibun), without person, gender, or number features, were thought to raise to INFL by successive cyclic head-to-head movement. Because only heads could undergo head movement, polymorphemic XP reflexives (e.g., himself) had to adjoin instead to XPs, specifically their immediate VPs. Therefore, whereas X⁰ reflexives could be bound long-distance, XP reflexives could be bound only locally. What accounted for monomorphemic reflexives’ subject orientation was the idea that, raised to INFL, they were c-commanded by subject NPs, but not by object NPs. In contrast, polymorphemic reflexives were c-commanded by non-subject NPs as well as by subject NPs when adjoined to VP. Thus, as Yuan (1998) puts it, long-distance binding and subject orientation were treated as “two coherent outcomes” of LF movement (p. 326).

Cole et al. (1990) also related antecedent types to a reflexive’s movement possibilities. An important contrast for Cole et al., however—similar to the distinction between
morphological and anaphoric AGR discussed above—was between functional and lexical INFL. In languages with functional INFL, like English, movement of a reflexive out of a VP was ruled out because it would cross a barrier, whereas in languages with lexical INFL, like Chinese, such movement was unproblematic.

Another LF-movement approach, by Katada (1988, 1989, 1991), is significant in the sense that it was designed specifically to account for the behavior of reflexives in Japanese, and also in the sense that certain of its assumptions were diametrically opposed to those just discussed. Katada divided reflexives into three types: non-operator anaphors (*kare-zisin*, ‘himself’), operator anaphors (*zibun*, ‘self’), and phrasal operator anaphors (*zibun-zisin*, ‘self-self’). The ability of *kare-zisin* to take both subject and non-subject antecedents, but only locally, was said to be related to the fact that it was c-commanded by both and to its possession of ϕ-features, which meant that there was no need for it to move to be interpreted. The operator anaphors, in contrast, lacking ϕ-features, were argued to raise into A-bar positions. *Zibun* was proposed to adjoin to VPs (possibly iteratively), where, according to Katada, it would be c-commanded by subjects, but not by non-subjects. *Zibun-zisin* was considered to be adjoined to the local VP, where it had to be bound by a local subject, because its *zibun* component lacked Case and could not move past where it would be governed. Interestingly, VP adjunction is precisely what Pica (1987), Battistella (1989), and Cole et al. (1990) proposed for polymorphemic reflexives and what they used to explain why such reflexives could be c-commanded by non-subjects—the opposite of what Katada assumed for monomorphemic *zibun* (Thomas, 1995).
Theoretical discrepancies such as these may not necessarily be cause for concern, especially if healthy debates can lead to more refined and accurate analyses. What seems possibly more worrying, though, is that researchers in this area appear to have been assuming and trying to explain a generalization about a relationship between morphology and binding behavior that does not actually hold cross-linguistically (as seen in the list of counter-examples reviewed above). Another current problem in this area of research, also alluded to above, is that recent developments in linguistic theory, advanced under the banner of the Minimalist Program, have “failed to adequately bring the binding theory into the fold” (Hicks, 2009, p. 6). Approaches to binding get “a particularly raw deal” in Minimalism because “even the most straightforward assumptions and technical devices adopted by previous grammatical models in order to capture binding facts become highly contentious when reexamined from a Minimalist perspective” (pp. 55, 11). For instance, in Minimalism, ad-hoc concepts (such as accessible SUBJECTs) are rejected as non-primitive notions not meeting ‘virtual conceptual necessity’ (p. 57), and it is not considered permissible to propose syntactic mechanisms or definitions of locality with the sole purpose of covering data, as was often done to account for binding phenomena under early approaches (p. 36).

In Hicks’ (2006) dissertation and (2009) book, he argues against viewing binding as taking place at LF. Instead, pointing out that Agree is “an operation to establish interpositional dependencies” (p. 40), he claims that binding relations are determined in narrow syntax (the computational component of the grammar) by the core operations of Agree and Merge. Constraints on binding such as c-command and locality, which were previously stipulated, are seen as natural outcomes of these more general mechanisms. Drawing on insights by Hornstein
Hicks endeavors to account for data previously seen as falling under Condition A by “reducing the behavior of reflexives... to their lexical feature specification” (p. 9). He proposes adding a new feature to the lexical entries of DPs in order to deal with the fact that anaphors are “obligatorily referentially dependent” (p. 98): Whereas pronouns and referential DPs would bear a valued referential feature, reflexives would bear an unvalued feature, and since features without values are uninterpretable, they would have to “enter into a syntactic dependency during the course of the derivation capable of valuing [them]” (p. 107). Agree is the only operation which can do so, and feature matching has to operate within a “computationally accessible domain” (p. 107) (the LF-phase, p. 126); thus, Agree would automatically serve to constrain locality within a domain that exists in Minimalist theory independently of binding concerns—“a long-time ‘holy grail’ of generative syntax” (p. 97). Hicks claims that this leads to “a complete elimination of Condition A from the grammar, without any loss of empirical coverage” and to an elimination of the binding theory as a component of UG (p. 98).

The point in mentioning these fundamental changes is not to suggest that SLA researchers might as well be atheoretical in their approaches to experimentation; theoretical frameworks can help to guide the formulation of research questions and the interpretation of results. They can also help to create well-thought-out rationales to underlie the selection of sentence types for use in experimental research. Even when theories change, the data which have been gathered from L2 learners remain. If these data are systematized in some way, they might contribute more readily to well-organized descriptions of L2 knowledge and aid researchers in forming clearer ideas about what learners tend to acquire and what they do not.
In the current study, with one of the goals being to test S. E. Carroll’s (2001) predictions about the potential effectiveness of feedback on linguistic phenomena for which relevant distinctions are not represented in surface forms, one important theoretical consideration is the possibility of overt morphological triggers of binding behavior. Given feedback researchers’ longstanding interest in the issue of negative evidence, another important consideration relates to the apparent role (or not) of L1 influence, which several empirical studies have explored.

*Previous empirical research on the second language acquisition of reflexives*

Empirical studies on the L2 acquisition of reflexives have usually focused on L1-L2 combinations in which one of the languages permits long-distance (LD) binding with a subject orientation whereas the other permits only local binding but without restrictions on the type of antecedent. The research questions have often involved issues related to Universal Grammar and the role of L1 influence, with researchers attempting to ascertain whether learners’ interlanguages show characteristics indicative of a ‘determined’ grammar (i.e., one they could have arrived at on the basis of L1 transfer or induction from L2 input) or an ‘underdetermined’ grammar (i.e., one apparently requiring information supplied by UG) (Hamilton, 1998).

These studies have produced a variety of results which, as will be reviewed below, have been interpreted as (1) evidence of learners following universal principles, such as sensitivity to c-command (e.g., Thomas, 1991; White, 1995b); (2) a lack of evidence that adult L2 learners follow {M,W}’s Subset Principle (e.g., Hirakawa, 1990; Thomas 1989); (3) possible evidence against a necessary relationship between a subject orientation and long-distance binding of reflexives in L2 grammars (e.g., Thomas, 1995; see also Christie, 1992; Wakabayashi, 1996); (4)
some evidence for L1 influence, at least initially (e.g., Hirakawa, 1990; MacLaughlin, 1998; Matsumura, 1994; Yuan, 1998); and also (5) patterns that do not seem to be explainable with reference to L1 influence (e.g., Finer & Broselow, 1986; Thomas, 1989, 1991; White, 1995b). Researchers have also explored and theorized about relationships between the morphological features of reflexives and their binding domains (e.g., Akiyama, 2002; L. Jiang, 2009; Wells, 1998; Yip & Tang, 1998) and about whether explicit instruction on one characteristic of reflexives could lead to knowledge of another (White, 1995b; White et al., 1996).

**Results interpreted as evidence of learners following universal principles**

To investigate the issue of learners’ success with universal constraints on the interpretation of reflexives, Thomas (1991) conducted a study with L1-Japanese and L1-Spanish learners of L2 English, and with L1-English and L1-Chinese learners of L2 Japanese. In addition to including sentences that would allow her to test for a subject orientation and long-distance antecedents (with respect to which certain Chinese and Japanese reflexives differed from English and Spanish ones), she also included sentences that would allow her to test for universally disallowed non-c-commanding antecedents and strategies based on linear order as opposed to structural properties, such as *Mika-no imooto-ga zibun-no mondai-nituite hanasita* (‘Mika’s sister spoke about self’s problems’) and *The man who John met wrote a story about himself* (p. 222). The idea was that if learners allowed Mika or John as antecedents of the reflexives in these sentences, it would indicate “structure-independent” behavior going against universal principles since neither c-commanded the reflexive.
For the Japanese sentences with possessives, Thomas found that the majority of L2 learners chose the c-commanding antecedent (imooto, ‘sister’); none consistently selected only the non-c-commanding option (Mika), and only 3 out of 41 (7%) consistently interpreted sentences of that type as ambiguous (wrongly allowing either Mika or imooto to serve as the antecedent of zibun). For the English sentences with relative clauses, 60-96% of the L2 learners (across proficiency levels) consistently chose the c-commanding antecedent (man), and they did not seem to employ linear-order strategies identifying the reflexive with the closest, non-c-commanding NP (John); only 2.7% did that consistently. Thomas (1991) interprets these results as indicating that L2 learners “rarely form hypotheses not sanctioned by UG” (p. 230).

Under the (inaccurate) assumption that no natural languages display a binding pattern disallowing subjects as antecedents, Thomas also inspected whether any individual learners systematically bound reflexives exclusively to non-subjects or exclusively to LD antecedents. Both Finer (1991) and Hirakawa (1990) had found that about 20% of their participants’ responses in L2 English bound reflexives only to non-subjects like Mr. Thin in monoclausal sentences like Mr. Fat gives Mr. Thin a picture of himself, and Thomas wondered whether this might be evidence of non-UG-sanctioned behavior. In her study, however, she found that only 4 out of 132 participants (3%) bound reflexives exclusively to non-subjects in L2 English, and none did so in L2 Japanese. Also, fewer than 3% bound reflexives exclusively to LD antecedents in L2 English. Accordingly, Thomas concluded that L2 learners tend to observe UG constraints. One “discordant note” was that 50% of the L1-Chinese speakers consistently bound reflexives only

---

12 As mentioned above, according to Hamilton (1998, referring to a personal communication with Veturlídi Oskarsson, 1996), hann sjálfur in Icelandic takes only object antecedents, and other Scandinavian languages may also have “anti-subject-oriented” reflexives (p. 296).
to LD antecedents in L2 Japanese even though local antecedents should also have been allowed. Since the L1-Japanese participants showed a similar tendency in Japanese, however, she reasoned that it might simply reflect a preference or amount to a side-effect of the task design.

**Results interpreted as evidence against the Subset Principle**

In an earlier study, Thomas (1989) conducted an experiment with 97 learners of L2 English to investigate whether they showed a preference for subjects as antecedents (a ‘subject strategy’, SS) and whether they obeyed a ‘clause-mate condition’ (CMC). The latter was to test Manzini and Wexler’s Governing Category Parameter in adult SLA. The participants ranged in proficiency from low-intermediate to advanced and came from 20 different L1 backgrounds (with 29 L1-Spanish and 24 L1-Chinese speakers comprising 2 major subsets). Thomas also tested 11 native speakers of English and 4 participants she considered bilingual.\(^{13}\) She employed 4 sentence types: 2 biclausal types to test the CMC and 2 monoclausal types to test for an SS. In each of these categories, one sentence type was designed to be biased in such a way that the disallowed or non-subject option would seem more plausible (e.g., ‘Mary’ in *Mary angrily told me that Sue had spilled a lot of paint on herself*), whereas the other sentence type was designed to be neutral.

Thomas (1989) found that, similarly to native speakers, the L2 learners seemed to show a preference for subject antecedents: Most were (rightly) willing to accept the non-subject NP when it was favored in biased monoclausal sentences, but otherwise (i.e., in neutral sentences)

---

\(^{13}\) One of the “native speakers” of English was actually a composite of 20 people, 10 of whom took the equivalent of one half of the test, and 10 of whom took the equivalent of the other half, as distractor items in another experiment (Thomas, 1989, p. 300, note 12).
they tended to select the subject NP. The CMC, however, was “not fully captured” by the L2 learners (p. 290); unlike native speakers of English, the learners selected exclusively the clause-mate antecedent in only 65% of neutral biclausal sentences and 49% of biased biclausal sentences (a “low incidence of clause mate binding,” according to Thomas, p. 297). In other words, contra {M,W}'s Subset Principle, they “set the governing category parameter too widely without positive evidence” (p. 281). As will be discussed in greater detail below in relation to the question of transfer, Thomas found little difference between the L1-Chinese and L1-Spanish speakers despite the fact that Chinese allows long-distance binding whereas Spanish does not.

A study by Hirakawa (1990), conducted with Japanese-speaking learners of L2 English, also sought to evaluate the applicability of {M,W}'s proposed parameters in adult L2 acquisition. Her interest was in asking whether learners would obey the Subset Principle, starting out with the most restrictive value of the GCP; whether they would transfer a less restricted L1 value, allowing LD antecedents; whether they would show behavior distinct from the L1 or L2, perhaps assuming an intermediate value of the GCP; or whether they would show a supposedly unsanctioned grammar, such as one allowing only nonlocal or non-subject antecedents. The participants included 4 groups of L2-English learners (age 15-19, in grades 10-13), most of whom had started learning English at age 12; a group of L1-Japanese speakers (age 17-18, in grade 12); and a group of L1-English speakers (age 17-19, in college). Hirakawa used 5 sentence types: biclausal finite and nonfinite, triclausal finite and nonfinite, and monoclausal with an indirect object.

Examining the results for the multiclausal sentences, Hirakawa found that although learners most frequently selected the correct (local) responses, there were also participants
who allowed long-distance antecedents for all 4 multiclausal sentence types, with the incidence of LD binding ranging from 17-36%. From Hirakawa’s perspective, “the extent of non-local responses [was] quite large” (p. 71), and since the participants apparently “did not start from the most unmarked, smallest grammar,” she interprets the results as evidence that L2 learners make transfer errors and that “the Subset Principle does not operate in L2 acquisition” (p. 76).

In contrast to Thomas (1989) and Hirakawa (1990), Wakabayashi (1996) focused more on the ordering of values in the GCP and PAP than on the strict applicability of the Subset Principle, and he interpreted his results (as well as a re-analysis of Hirakawa’s results) as suggesting that \{M,W\}’s parameters operate in SLA. His participants were 40 Japanese-speaking learners of L2 English who, like Hirakawa’s, had begun studying English around age 12 and had been learning it for at least 6 years by the time of the experiment. The controls were 21 native speakers of English. Wakabayashi employed 1 sentence type to test the PAP (monoclausal with a non-subject antecedent) and 4 sentence types to test the GCP (biclausal finite, biclausal nonfinite, monoclausal with 1 possible antecedent, and monoclausal with an antecedent in a complex NP [e.g., Yosuke read Wataru’s criticism about himself]). He analyzed the participants’ responses in two ways: with regard to their first choices and with regard to their reports of all possible antecedents. He also analyzed the results in two ways: in terms of group averages and at the level of individual participants who answered consistently in a certain way. Wakabayashi reasoned that if a learner allowed antecedents at the longest distance while not allowing antecedents at a more restricted intermediate distance, that would suggest “some system which is not under the sanction of GCP” (p. 275).
From all four perspectives of analysis (i.e., in terms of group averages, individual participants, first choices, and all allowed antecedents), Wakabayashi found that the results followed the predicted order of the GCP. For example, the percentages of participants showing targetlike behavior with regard to all allowed antecedents were 95% for monoclausal single-antecedent sentences, 63% for biclausal finite, 40% for biclausal nonfinite, and 23% for monoclausal complex-NP sentences. Also, at the individual level, when a participant restricted antecedents appropriately in a more-restricted sentence type (e.g., with complex NPs), s/he would also restrict antecedents appropriately in the less-restricted type(s) (e.g., biclausal nonfinite) (p. 283). Two participants were exceptional in that they consistently accepted any antecedents in all sentences, but no participants consistently allowed further-away antecedents while rejecting closer ones.

Regarding the PAP, both the L2 learners and the native speakers tended to answer that only subject antecedents were possible, despite the fact that both subjects and non-subjects are acceptable in English. This was whether their responses were analyzed with regard to their first choices (84% only subjects for L2 learners, 68% for NSs) or with regard to their reports of all allowable antecedents (71% only subjects for L2 learners, 54% for NSs). Thus, as in Thomas’ (1989) study, it is possible that “the experiment did not succeed in revealing the participants’ [or native speakers’] competence” (Wakabayashi, 1996, p. 281).

In re-analyzing Hirakawa’s (1990) data, Wakabayashi was able to classify her participants as belonging to 3 groups: (1) those consistently selecting local antecedents in both finite and nonfinite sentences, (2) those consistently selecting local antecedents in finite sentences, but not in nonfinite sentences, and (3) those not consistently selecting local
antecedents in either finite or nonfinite sentences. Wakabayashi acknowledges that “[s]ome of Hirakawa’s subjects seem[ed] consistently to interpret that reflexives exclusively refer to long-distance NPs” (p. 290, note 25), in apparent violation of the GCP. However, reasoning that this might simply reflect first-choice preferences, he maintains his conclusion that L2 learners’ grammars tend to be compatible with the order of {M,W}’s parametric values. In other words, learners may not start out assuming the most restricted values of parameters in SLA, as the Subset Principle would have it, but something about the acquisition of reflexives may, in a sense, correspond to the predictions of the implicational hierarchy.

Long-distance binding and subject orientation

Once the Relativized SUBJECT and LF-movement approaches to explaining the behavior of reflexives had been proposed, some researchers sought to investigate their applicability to adult SLA by looking for relationships between subject orientation and long-distance binding in L2 learners’ grammars (e.g., Christie, 1992; Thomas, 1995; Wakabayashi, 1996; Yuan, 1998). In one such study, Christie (1992) used cluster analysis to assess the degree of correlation between subject orientation and LD binding, as demonstrated by L2 learners’ responses on a picture-identification task. The participants were L1-Spanish and L1-Chinese learners of L2 English, L1-English learners of L2 Spanish, and L1-English learners of L2 Chinese. Not finding evidence for such a relationship, Christie concluded that her data did not support LF-movement approaches or provide evidence of access to UG by L2 learners. Wakabayashi (1996) also failed to find a correlation between binding distance and subject orientation. Moreover, looking at the binding patterns of individual participants, he determined that 32% (12/38) violated the
assumed relationship in the sense that their responses exhibited LD binding without also exhibiting a subject orientation.

A problem with both studies, according to Thomas (1995), is that the Relativized SUBJECT and LF-movement approaches do not predict general correlations between LD binding and subject orientation; rather, they predict implicational relationships: LD binding should entail that the antecedent is a subject, but not vice versa, and non-subject antecedents should entail local binding, but not vice versa (p. 213). In fact, the native-speaker controls in Christie’s study did not show the correlation she had predicted, either. In order to be able to argue either for or against the theoretical approaches in question, Thomas claims, it is necessary to examine learners’ interpretations of reflexives in biclausal sentences with LD non-subjects proposed as antecedents.14

Pursuing this idea, as an initial step, Thomas (1995) re-analyzed data collected from 34 L1-English learners of L2 Japanese (previously analyzed in her own [1991, 1993] studies) to see whether the learners who allowed LD binding for zibun also required subject antecedents for it. (Her earlier dataset did not include the crucial LD non-subject antecedents.) In doing so, she found that 28 of the 34 learners appeared to classify zibun as a reflexive requiring c-commanding antecedents, 7 of those 28 consistently allowed LD binding, and 5 of those 7 consistently required subject antecedents (“consistent” meaning 3 out of 4 times). Most relevant to her research question were the 2/7 learners with LD binding who did not require subject antecedents. Thomas acknowledges that this last set might “have attributed to zibun some of the properties instantiated in... Serbo-Croatian sebe” (p. 220), which allows LD subjects

14 Wakabayashi acknowledges this and states that his results may be inconclusive because of it (p. 294, note 38).
in addition to local subjects and objects. While noting that it is not clear how LF-movement theories might be able to account for this, she does not claim based on these results to have found evidence of learner grammars that violate UG, considering that this is a pattern attested in natural language.\(^{15}\)

The new participants in Thomas’ (1995) study were 58 adult learners of L2 Japanese from a variety of L1 backgrounds: 32 English; 20 Chinese, Korean, or Thai (languages which, like Japanese, have LD reflexives); and 6 French, Spanish, or German. A set of lower-proficiency learners (n=34) were enrolled in third-semester Japanese classes, while a set of higher-proficiency learners were enrolled in either fourth-year courses (n=5) or an intensive course in advanced technical Japanese (n=19). Thomas also included 34 native speakers of Japanese as a control group. The 4 sentence types employed in her study, with 4 tokens of each in a story- and picture-based truth-value judgment task, were biclausal sentences testing acceptance of LD subjects, biclausals testing (crucially) rejection of LD non-subjects, monoclausal sentences testing acceptance of local subjects, and monoclausal sentences testing rejection of local non-subjects.

In group-based analyses, Thomas found the following: On the biclausal sentences testing acceptance of LD subjects, both the low- and high-proficiency learners performed only at around chance levels; on the biclausals testing rejection of LD non-subjects, the high-proficiency learners performed like native speakers, but the low-proficiency learners again performed at around chance; on monoclausal sentences testing acceptance of local subjects, both the low- and high-proficiency groups appeared to perform like native speakers; and on monoclausal sentences testing rejection of local non-subjects, the high-proficiency learners performed

\(^{15}\) Yuan (1998) also found that a few L1-English and L1-Japanese speakers allowed both LD-subject and local-object binding in L2 Chinese, but he likewise notes that this combination violates only Chinese grammar, not UG (p. 336).
like native speakers, but the low-proficiency learners performed at around chance. In other words, the high-proficiency L2 learners mostly performed like native speakers, except that they allowed less LD binding to subject antecedents. The low-proficiency learners performed quite differently, however, tending to answer better than chance only with local subject antecedents.

Turning to the individual analyses, Thomas found that 54 of her 58 participants consistently allowed local subject antecedents in monoclusal sentences; 23 of those 54 consistently also allowed LD subjects in biclausals; and 16 of those 23 (who correctly allowed local and LD subject antecedents) also consistently rejected LD non-subjects in biclausals. (Incidentally, 12 of the 16 correctly banned local as well as LD non-subjects.) What this left was 7/23 learners (i.e., 30% of those Thomas considered relevant) who allowed LD binding of *zibun* to non-subjects and who therefore “may have grammars unaccounted for by UG if learners move *zibun* at LF” (p. 227). Thomas points out that this last group was mostly made up of low-proficiency learners. Only half (6 of 12) of the relevant subset of low-proficiency participants who allowed LD subjects also consistently and correctly rejected LD non-subjects; however, 91% (10 of 11) of the high-proficiency participants who allowed LD subjects consistently rejected LD non-subjects, and this latter group included learners both with and without L1s with long-distance reflexives.

Thomas interprets these results as indicating that “learners’ knowledge of *zibun* at a high-proficiency level is largely consistent with a key prediction of the movement in LF approach... [but] data from lower-proficiency learners are less readily accounted for” (p. 206). In light of this, although she admits the experiment was not designed to test the idea, she suggests that perhaps some of the 7 learners showing apparently ‘rogue’ grammars actually
treated *zibun* as a pronoun. Yuan (1998) rejects this interpretation, pointing out that 3 learners in his study similarly allowed LD subjects and LD non-subjects as antecedents of *ziji*, but they also allowed local subject binding, “which is clear evidence that *ziji* was treated as a reflexive rather than a pronoun” (p. 337). Alternative explanations are possible, of course. Hamilton (1998) suggests, for instance, that some of Thomas’ participants may have treated *zibun* as a logophor. It may also be relevant to consider that whereas positive evidence of LD antecedents may be available, negative evidence against non-subject antecedents (e.g., via correction or metalinguistic information) likely is not. Whatever the explanation, it seems not necessarily to be the case that L2 learners will obey a prohibition against LD non-subject antecedents.

**Results interpreted as evidence for first-language influence**

Kim, Montrul, and Yoon (2010) estimate that “when L2 binding properties are different from those of L1 grammars, transfer effects are usually shown in L2 acquisition” (p. 78). This may not be hedged sufficiently, considering the number of researchers who have argued that the results of their studies cannot be explained in terms of L1 influence, but it is true that learners whose L1s allow LD antecedents for reflexives have sometimes been found to perform more accurately with LD binding in L2s that also allow them, compared to learners whose L1s do not allow LD antecedents (e.g., Thomas, 1991; Yuan, 1998). Also, although it may not be warranted to interpret results as evidence of L1 influence when no comparisons have been made against learners whose L1s work differently, researchers conducting studies with participants of a single L1 have reported, for example, that some learners whose L1 allows LD antecedents will accept them (wrongly) in L2 English (e.g., Hirakawa, 1990; Matsumura, 1994),
and learners whose dominant language allows only local antecedents may tend not to accept LD antecedents in an L2 even when the L2 allows them (e.g., Kim, Montrul, & Yoon, 2009).

In Thomas’ (1991) experiment, in addition to including sentences to test for universal constraints (as discussed above), she included biclausal sentences with local and LD antecedents, and monoclausal sentences with subject and non-subject antecedents, to check whether learners from different L1 backgrounds would perform differently. As already explained, she conducted the study with L1-Japanese and L1-Spanish learners of L2 English, and also with L1-Chinese and L1-English learners of L2 Japanese. The results for L2 English will be reviewed later since they do not seem to suggest L1 influence. However, the results for L2 Japanese will be reviewed here since (in contrast to Thomas’ interpretations) they might.

Chinese *ziji* (like Japanese *zibun*) allows LD antecedents while disallowing non-subjects, whereas English *himself/herself* disallows LD antecedents while allowing non-subjects. Thus, Thomas used biclausal sentences in Japanese to check for L1 influence on learners’ acceptance of LD antecedents and used monoclausal sentences in Japanese to check for L1 influence on learners’ acceptance of non-subject antecedents.

The different groups of L2 learners in Thomas’ (1991) study showed different response patterns. For the biclausal sentences, 60% of the native speakers in her Japanese control group consistently chose only LD antecedents, 0% consistently chose only local, and merely 10% consistently allowed both. (This is despite the fact that both types are theoretically allowed. In other words, it is important to keep in mind that the task design does not seem to have avoided the problem of preferences.) Similarly to the L1-Japanese speakers, 50% of the learners who spoke Chinese as L1 consistently chose only LD antecedents in L2 Japanese, 25% consistently...
chose only local, and 0% consistently allowed both. The learners who spoke English as L1 were separated into low-, mid-, and high-proficiency groups and produced the following results: Only 1 low- and 1 high-proficiency learner consistently chose only LD antecedents in L2 Japanese; 38% of low-, 83% of mid-, and 23% of high-proficiency learners consistently chose only local; and 13% of low-, 8% of mid-, and 31% of high-proficiency learners consistently allowed both. For the monoclausal sentences, all of the native speakers of Japanese in the control group consistently required subjects as antecedents, rejecting non-subjects, as expected. Similarly, 88% of the L1-Chinese learners consistently required subjects in L2 Japanese. However, only 33-54% of the L1-English learners did so. While none of the L1-English learners consistently required non-subjects as antecedents in L2 Japanese, a number of them wrongly accepted either subjects or non-subjects.

Somewhat surprisingly, Thomas (1991) interprets her results overall as being inconsistent with the idea of L1 transfer, stating that “English- and Chinese-speaking learners of Japanese differ only in the frequency of long-distance responses in Type J1 [i.e., Japanese biclausal] (p<.0001), and subject responses in Type J3 [Japanese monoclausal with non-subject options] (p=.0093)” (p. 233, note 39). In view of the fact that there were only 3 sentence types in total, the other of which involved a theoretically universal c-command requirement, Thomas’ emphasis seems a bit misplaced. It may be true that learners “do not realize in L2 only the version of UG that is present in their native language” (p. 233), but these results regarding the acquisition of L2-Japanese reflexives by L1-English versus L1-Chinese learners do appear to indicate that L1 influence might play a role.
A study by Yuan (1998) was similar to Thomas’ (1991) study in that it investigated the L2 acquisition of a language with long-distance anaphors (Chinese) by learners whose L1s either did (Japanese) or did not (English) also have them. One of Yuan’s research questions was whether, in acquiring the binding characteristics of the LD Chinese reflexive *ziji*, L1-Japanese learners would show an advantage over L1-English learners. The participants were 57 adult English-speaking learners of Chinese, divided into 2 proficiency groups (intermediate and advanced), and 24 adult Japanese-speaking learners of Chinese, determined to be at an intermediate level of proficiency. A control group of 24 native speakers of Chinese was also included. Yuan followed Thomas (1989, 1991) in employing both neutral and biased versions of each sentence type, reasoning that if the learners truly observed syntactic constraints, “they should reject syntactically impossible coreference even if the pragmatics favor it” (p. 329).

There were 11 kinds of sentences: 5 biclausal finite (neutral between LD and local subjects, neutral between LD subjects and LD non-subjects, favoring a LD subject antecedent, favoring a LD non-subject antecedent, and favoring a local subject antecedent), 3 biclausal nonfinite (neutral, favoring LD, and favoring local), and 3 monoclausal (neutral, favoring a subject antecedent, and favoring a non-subject antecedent).

Before reviewing Yuan’s results, it is important to point out that he collapsed the data for certain sentence types for certain analyses and looked only at pragmatically biased sentence types for other analyses. For example, in investigating learners’ acceptance of LD binding, he reported statistical analyses only for the finite and nonfinite biclausal sentences in which the LD antecedents had been favored. In doing so, he found that the L1-Japanese participants correctly allowed more LD binding in L2 Chinese than the L1-English groups did. Moreover, whereas
there was no statistically significant difference between the L1-Japanese learners and the native speakers of Chinese on those sentence types, some differences did emerge between the L1-English learners and the Chinese NSs (namely, on finite sentences for both proficiency groups and on nonfinite sentences for the intermediate-level learners).

In investigating whether the various groups of participants correctly disallowed non-subject antecedents in Chinese, Yuan likewise looked at the sentences where those options had been favored and found that subject orientation seemed to be problematic for all learners. On the monoclausal sentences, whereas the native speakers of Chinese accepted local non-subjects (experimentally favored, but disallowed) only 6% of the time, the L1-Japanese learners accepted them at a rate of 49%, and the L1-English learners accepted them at rates of 52% (intermediate) and 32% (advanced). On the biclausal sentences, whereas the NSs accepted LD non-subjects (also experimentally favored, but disallowed) only 3% of the time, the L1-Japanese learners accepted them at a rate of 31%, and the L1-English learners accepted them at rates of 50% (intermediate) and 28% (advanced).

In interpreting these results, Yuan (1998) argues that L1 influence does occur, but he notes that not everything can be explained in that way: Evidence for L1 effects can be inferred from the finding that L1-Japanese speakers appear to outperform L1-English speakers with regard to LD binding in L2 Chinese; however, it cannot be inferred from the results for subject orientation because both L1-Japanese and L1-English speakers had trouble rejecting non-subject antecedents in L2 Chinese. According to Yuan, the former set of findings (regarding LD anaphora) may have been obtained because “Japanese learners recognize ziji as a counterpart of zibun and are exposed to data indicating the property of LD binding of ziji” (p. 335). He
considers it “likely that English learners initially transfer [local] binding” and this interference from the L1 “delays [their] acquisition of LD binding of ziji” (p. 335).

Under this sort of analysis, it is not completely clear why L1-Japanese learners who hypothetically identified ziji with zibun would not also transfer zibun’s subject orientation, but this appears to be what Yuan argues when he claims that L1 influence is not involved in the latter set of findings. He reasons that L1-Japanese speakers might experience difficulties with subject orientation similar to those experienced by L1-English speakers due to “misleading evidence in the input data” (p. 336). After explaining that ziji can sometimes be bound by PRO or null subjects which are controlled by or co-indexed with objects (e.g., in Mama gaosu nuer, PRO, buyao jiaoguan ziji, ‘The mother told the daughter not to spoil herself’), he states that “the evidence at surface structure... is likely to give wrong signals that ziji in Chinese can be object-bound and therefore has a property of free orientation” (p. 336). Although object control also exists in Japanese, Yuan argues that “the similar structure in their L1 cannot guarantee that Japanese learners would analyze Chinese data... correctly” (p. 336, note 5). Without independent evidence regarding learners’ thought processes (e.g., from introspective data produced in think-aloud protocols), it seems somewhat ad-hoc for Yuan to invoke the possibility that L1 influence would not occur and inappropriate (metalinguistic?) analyses would be triggered by surface features in specifically these cases. Even without a compelling explanation, though, the empirical results stand: L2 performance sometimes, but not always, appears to be influenced by L1 background, and situations involving positive evidence may be more straightforward for acquisition.
Results interpreted as evidence against first-language influence

In contrast to certain results of Thomas’ (1991) and Yuan’s (1998) studies reviewed above, other experimental findings have not seemed explicable in terms of L1 influence. They can be classified into at least two major categories: those in which properties of the L1 have not appeared to influence L2 binding patterns (e.g., Thomas, 1989, 1991, 1995; White, 1995b; White et al., 1996) and those in which learners have shown a particular binding pattern (often referred to as a “finite/nonfinite asymmetry”) found in neither the L1 nor the L2 under investigation (e.g., Akiyama, 2002; Finer, 1991; Finer & Broselow, 1986; Hirakawa, 1990; MacLaughlin, 1998; Matusumura, 1994; Wakabayashi, 1996; White, 1995b; Yuan, 1998). In the case of the finite/nonfinite asymmetry, though, it has been argued that L1 influence might be a factor in a more complex way (L. Jiang, 2009).

No apparent L1 influence on L2 binding patterns

Details on the methods of Thomas’ (1989, 1991) studies have already been given and so will not be repeated here. In the 1989 experiment, investigating whether learners followed a clause-mate condition and subject strategy, Thomas found that both L1-Spanish and L1-Chinese learners of L2 English wrongly allowed long-distance binding “at rates much higher than [did] native speakers” (p. 291). This was despite the generalization that, like English himself/herself, Spanish reflexives must also be co-referential with clause-mate antecedents (i.e., local), whereas Chinese has both local (taziji, ‘himself’) and LD (ziji, ‘self’) reflexives. Furthermore, similarly to the native speakers, both groups of learners showed a preference for subject antecedents over non-subjects in neutral sentences, while allowing non-subject antecedents in
pragmatically biased ones. Thomas (1989) thus found “little support for the prediction that the Spanish-speaking group [would] perform better because their L1 system... [was] closer to that of the L2 system” and concludes that “[a] transfer hypothesis cannot explain these results” (p. 291). In an endnote, Thomas acknowledges that data from 4 L1-Japanese and 8 L1-Korean speakers “may be evidence of transfer” (p. 300, note 15) since they produced higher rates of LD binding in neutral biclausal sentences than any of the other groups. However, Thomas notes that these learners also allowed non-subject antecedents in monoclausal sentences and argues that those results do not support a hypothesis of L1 influence.

In Thomas’ later (1991) experiment, as we have seen, the results for the L1-English and L1-Chinese learners of L2 Japanese might be interpretable as indicating L1 influence; however, the results for her L1-Japanese and L1-Spanish learners of L2 English do not seem to be. In biclausal English sentences, where LD binding is not allowed, 70-91% of every learner group (and 100% of native speakers) consistently chose only local antecedents, with the L1-Japanese and L1-Spanish learners behaving similarly: Only 5-20% of the Spanish speakers consistently reported that such sentences were ambiguous in English, and the vast majority of Japanese speakers also bound locally, with no statistically significant differences between the groups; no one consistently chose LD antecedents. In monoclausal sentences, where both subject and non-subject antecedents are allowed in English, most learners (50-81%) consistently chose only subject antecedents and a few (4-15%) consistently chose only non-subjects, but some (5-24%) did correctly allow both subjects and non-subjects. In this last category, all groups except the low-proficiency L1-Japanese learners showed a rate of around 20%.
Finally, in a set of studies involving attempts at instructional intervention, to be discussed in greater detail below, White (1995b) found that both L1-French and L1-Japanese learners of L2 English wrongly allowed LD binding, and White et al. (1996) found no differences between learners of L2 Japanese with French, English, Korean, and Chinese L1 backgrounds.

The finite/nonfinite asymmetry

The so-called “finite/nonfinite asymmetry” (also known as the “tensed/infinitive asymmetry”) has commonly—though wrongly, in Matsumura’s (2007) view—been accepted as “one of the basic facts of second language acquisition” (p. 323). It was first reported in a very small-scale experiment by Finer and Broselow (1986), who found that L1-Korean learners of L2 English showed higher rates of (nontargetlike) LD binding in biclausal sentences with nonfinite embedded clauses (e.g., Mr. Fat wants Mr. Thin PRO to paint himself) than in sentences with finite embedded clauses (e.g., Mr. Fat thinks that Mr. Thin will paint himself). Such an asymmetry does, in fact, exist in certain languages (e.g., Russian, whose reflexive sebjja can refer to an antecedent across the subject of a nonfinite clause, but not across the subject of a finite clause). However, crucially, it exists in neither English nor Korean. Thus, this phenomenon originally attracted a great deal of interest in generative SLA research in relation to the notion that, if neither the L1 nor the L2 being investigated exhibited the asymmetry, the only possible explanation for its appearance in interlanguage grammars would be innate linguistic knowledge.

At first, the finite/nonfinite asymmetry was interpreted with reference to {M,W}’s Governing Category Parameter, which, as explained above, had 5 values, the third of which (the Russian value, Type ‘c’) seemed to be in evidence. Later, researchers operating within an LF-
movement or Relativized SUBJECT framework focused on the morphological features of reflexives and verbs in order to explain it. Noting that Russian has morphological AGR along with its monomorphemic reflexive, they reasoned that if L2 learners (of English, for example) misanalyzed a (necessarily local) polymorphemic reflexive as monomorphemic (and therefore potentially long-distance), while correctly recognizing the verbal morphology in finite clauses, that would explain why they allowed LD antecedents outside of nonfinite clauses only (L. Jiang, 2009; White, 1995a). In other words, L1 influence could actually be involved, but in a complex way related to the characteristics of verbal and reflexive morphology in learners’ interlanguages.

To review the empirical results of some early studies briefly, Finer and Broselow (1986) found that their 6 L1-Korean, L2-English participants mostly interpreted reflexives in biclausal sentences with finite embedded clauses as referring to local antecedents, opting for LD ones only 8% of the time (and local ones in the other 92% of cases). In biclausal sentences with nonfinite embedded clauses, however, the rates were 58% local, 38% long-distance, and 4% ambiguous between local and long-distance. In a larger-scale study by Finer (1991), with 20 L1-Japanese and 30 L1-Korean learners of L2 English, much lower rates of LD binding and much smaller differences between finite and nonfinite clauses were found; however, there was still a statistically significant difference between the two clause types: L1-Japanese learners allowed LD antecedents in 7% of nonfinite and 2% of finite clauses, and L1-Korean learners allowed them in 12% of nonfinite and 5% of finite clauses. In Hirakawa’s (1990) experiment with L1-Japanese learners of English, the participants “made a considerably larger number of mistakes” than Finer and Broselow’s participants had with finite biclausal, allowing LD antecedents 23%
of the time (p. 77), but they were nonetheless significantly more accurate with finite biclausals than with nonfinite biclausals, for which they allowed LD antecedents 44% of the time.

Some later studies in which finite/nonfinite asymmetries were interpreted with reference to morphological characteristics include Yuan (1998), Akiyama (2002), and L. Jiang (2009). In Yuan’s experiment, already discussed above, an apparent asymmetry was reported between finite and nonfinite clauses for L1-English learners of L2 Chinese, but not for L1-Japanese learners of L2 Chinese: When the LD antecedent was pragmatically favored, the English speakers permitted LD binding significantly more often across nonfinite clauses (69-88% of the time) than across finite ones (53-71%). However, though Yuan does not remark on it, this does not appear to have been the case for the pragmatically neutral sentences, and, in fact, the finite/nonfinite asymmetry may even have trended in the opposite direction for those items, with LD binding seeming to be more admissible across finite clauses (40-48%) than across nonfinite ones (15-35%). Moreover, the pragmatic favoring of LD antecedents may have inadvertently been stronger in Yuan’s biclausal nonfinite sentences (e.g., nonfinite: *Gao Hong bu yuanyi gen bieren jianghua, suoyi ta bu yuanyi ta mama xiang bieren jieshao ziji* [Gao Hong does not like to talk to other people; therefore, she does not like her mother to introduce self to other people], versus finite: *Wang Ming bu gaoxing de shuo Li Dong jingchang bu xiangxinziji* [Wang Ming said unhappily that Li Dong often does not trust self]), potentially contributing to the finding of an asymmetry. In relation to Yuan’s idea (mentioned above) that learners might have been interpreting the embedded subjects in nonfinite clauses as local objects (in which case those sentences would count as monoclausal and the subjects would not count as long-distance), other researchers have argued that learners assuming such structures would be
“operating outside the constraints imposed by UG” since the Projection Principle, Theta Criterion, and X-bar Theory (“or their current theoretical equivalents”) require a subject to be projected for an embedded verb (MacLaughlin, 1998, p. 207, see also Wakabayashi, 1996). In a word, there may be cause for skepticism regarding the finite/nonfinite asymmetry reported in Yuan’s (1998) experiment.

Akiyama’s (2002) results indicated a finite/nonfinite asymmetry more clearly; however, skeptical about the adequacy of previous theories to explain his results, he avoided speculating about potential explanations. Instead, he stated simply that, although his study “clearly show[ed] that the tensed-infinitive asymmetry exists throughout all proficiency levels in the case of Japanese learners... [and] the locality condition on English reflexives develops differently depending on clause type” (p. 43), “the UG models proposed thus far fail to account for the L2 learning manifested by the findings” (p. 45). Akiyama’s participants were 285 L1-Japanese learners of L2 English judged to represent 5 proficiency levels according to the amount of time they had been studying. They were junior- and senior-high-school students in grades 8-11 in Japan and university students (mostly pursuing graduate degrees) in the US. There were also 20 L1-English and 20 L1-Japanese native speakers who served as control groups. Only 3 sentence types were included: biclausal with a finite embedded clause (3 tokens), biclausal with a nonfinite embedded clause (3 tokens, using only the verb want), and monoclausal with a prepositional phrase to check for sensitivity to c-command (2 tokens, considered distractors).

In group-based analyses, Akiyama found that the L2 learners consistently performed better on finite items than on nonfinite items across all proficiency levels, and development
was not observed for the items involving the rejection of LD antecedents across nonfinite clauses. In analyses based on the patterns shown consistently by individual learners (who were classified as [1] always restricting to local, [2] always allowing LD, [3] restricting to local for finite clauses but allowing LD for nonfinite clauses, or [4] contrary to theory, restricting to local for nonfinite but allowing LD for finite), Akiyama also found that the finite and nonfinite sentence types showed distinct patterns of development and that “there was an almost invariable ratio (around 30%) of participants at every level who allowed both local and LD binding for [nonfinite] but only local binding for [finite]” (p. 42). About half of the most advanced participants consistently required local antecedents for both clause types, but the fact that “an appreciable percentage” of them (about 35%) failed to reject LD antecedents seemed to Akiyama “extremely difficult to account for” (p. 27).

Akiyama points out that Progovac's (1992, 1993) approach would account for the asymmetry straightforwardly if the learners could be taken as interpreting polymorphemic *himself* as morphologically simple. In that case, an AGR-chain could be established across nonfinite clauses in English, but not across finite clauses, since infinitives lack overt agreement whereas finite verbs mark it morphologically. At the same time, though, Akiyama notes that this line of reasoning “crucially depends on the assumption that those who exhibit the tensed-infinitive asymmetry take English reflexives to be monomorphemic” (p. 45). In the context of this study, that would mean that the 35% of the advanced participants who showed the asymmetry would have to be assumed to misinterpret *himself* as monomorphemic, which, for Akiyama, “is very hard to imagine” (p. 45).
A study by L. Jiang (2009), similar to the experiment by Akiyama just described, also set out to investigate developmental patterns in the acquisition of locality constraints for reflexives in L2 English, and he interprets the results as suggesting essentially what Akiyama considered unlikely. The participants were 66 L1-Chinese learners of English, divided into 3 proficiency levels (beginning, intermediate, and advanced) on the basis of a placement test, and 12 native speakers of English serving as controls. The learners included 21 middle-school students and 45 university students learning English in China. Four biclausal sentence types were employed, with 6 tokens of each: (1) finite with a context suggesting a local antecedent, (2) finite with a context suggesting a long-distance antecedent, (3) nonfinite with a context suggesting local, and (4) nonfinite with a context suggesting long-distance.

In group-based analyses, Jiang found that learners “readily accepted local antecedents” in those sentences where the context supported them (p. 479), but that there were statistically significant differences among the groups (beginning, intermediate, advanced, and control) on both types of sentences (finite and nonfinite) when the context suggested a LD antecedent. The intermediate-level group appeared to be more sensitive than the others to the difference in clause types; they were significantly more accurate at rejecting LD antecedents across finite clauses than across nonfinite ones (p. 480).

In order to perform individual-based analyses, Jiang then classified each participant as consistently exhibiting one of three binding patterns (English-, Russian-, or Chinese-like) or as answering inconsistently. He found that learners at higher proficiency levels were more likely to show an English-like (local-only) pattern and less likely to show a Chinese-like (possibly LD) pattern. However, interestingly, the Russian-like pattern appeared to be most common for the
intermediate participants; it was much less common for the beginning and advanced students (p. 481), as also suggested by the group results.

Jiang interprets these findings with reference to Pica (1987) and Battistella’s (1989) LF-movement approach, explaining that beginning L1-Chinese learners of English may “misanalyze” himself as monomorphemic at first (“due to transfer of L1 knowledge of the monomorphemic reflexive ziji”) and also be “unaware” of tense markings in English finite clauses (“transfer[ring] the properties of tense in their L1 Chinese into their L2 English”); thus, for beginning learners, reflexives move to INFL across both finite and nonfinite clauses, and no asymmetry is observed (pp. 483-4). As the learners become more proficient in English (at an intermediate level), they become more sensitive to the tense marking in finite clauses, but they still misanalyze himself as monomorphemic, so reflexives are able to move to INFL across nonfinite clauses but not finite ones, and a finite/nonfinite asymmetry is found. By the time learners have reached an advanced level of proficiency, they are not only sensitive to verb morphology but also “realize” that English reflexives are polymorphemic; as such, they allow only local antecedents and again show no finite/nonfinite asymmetry, albeit for different reasons (p. 485). Although Jiang did not perform independent analyses of the learners’ knowledge of morphology, his transfer-based explanation at least potentially fits with the theories he adopted as a foundation for his study.16

16 Another researcher to suggest that L1 influence might explain a finite/nonfinite asymmetry was Bennett (1994). Her participants were native speakers of Serbo-Croatian learning English as an L2. Serbo-Croatian has overt verbal agreement and a monomorphemic reflexive which shows the asymmetry. When Bennett found that the learners allowed LD binding only out of nonfinite clauses in L2 English, she proposed that perhaps they treated English like their L1: rightly as having overt morphological agreement, but wrongly as having monomorphemic reflexives. Because of the direct similarity between characteristics of the L1 and the learners’ behavior in the L2, the study is not used to support arguments regarding innate universal knowledge. However, it might indirectly support Jiang’s argument about L1 influence. Unfortunately, as in the other studies discussed in this section, the learners’ knowledge of morphology was not assessed independently.
The researchers whose work has been discussed so far have differed rather strikingly in their assessments of how likely it is that their results can be explained in terms of proposed relationships between morphology and binding behavior. What Yuan (1998), Akiyama (2002), and L. Jiang (2009) have in common, though, is that none of them examined learners’ knowledge of morphology directly and, consequently, were constrained to speculating about it. It is therefore important to examine the results of studies designed specifically to investigate the relationships between various morphological and syntactic aspects of binding patterns.

**Long-distance binding and morphology**

A collection of research on the topic of morphology and its interfaces, edited by Beck (1998), includes reports of studies designed specifically to explore relationships between morphological characteristics and binding behavior in learners’ L2 grammars (e.g., Wells, 1998; Yip & Tang, 1998). In a study employing Progovac’s (1992, 1993) Relativized SUBJECT approach as a theoretical framework, Wells (1998) used multiple measures to test not only her participants’ L2 binding behavior, but also the morphological characteristics of both verbs and reflexives in their L2 grammars. The measures included an oral translation task to test for overt subject-verb agreement, a grammaticality judgment task (GJT) to test the multimorphemic nature of English reflexives (checking whether the learners would reject forms like “yesterday’s herself” since XP reflexives cannot be modified), and a story-based truth-value judgment task (TVJT) to test for local binding. Wells notes that Progovac and Connell (1991) had previously found L1-Chinese learners of L2 English to allow LD readings for anaphors like *herself* while allowing only local readings for forms like *her usual self*, where multiple morphemes would
seem clearly to be involved. However, the relationship between learners’ knowledge of verb morphology and binding remained to be investigated.

The participants in Wells’ (1998) study were 31 L1-Chinese learners of L2 English and a set of native speakers of English whom she included as controls. The TVJT included two relevant item types: 10 biclausal sentences with finite embedded clauses and 5 biclausal sentences with nonfinite embedded clauses. For each measure, Wells set a threshold above which she would count the learners as having demonstrated acquisition: 70% correct agreement markings on the oral translation test, 75% correct rejection of ungrammatical items on the GJT, and 80% correct rejection of LD antecedents for each of the two types of sentences on the binding test. Her stated hypotheses were that learners who had acquired neither the verbal nor the reflexive morphology would allow LD binding across both finite and nonfinite clauses; learners who had acquired only the verbal morphology would allow LD binding only across nonfinite clauses; and learners who had acquired the reflexive morphology would allow only local binding.

As it turned out, most of the participants demonstrated having acquired the morphological characteristics of both verbs and reflexives according to Wells’ criteria, and most usually showed local binding. They showed “somewhat more success” on the finite biclausals in that 27 of the 31 consistently rejected LD antecedents across finite clauses, but only 19 of the 31 consistently did so across nonfinite clauses (p. 242). To evaluate her hypotheses, Wells examined individual learners who showed particular combinations of morphological and binding characteristics. There were 6 learners who did not demonstrate having acquired verbal or reflexive morphology in English, and who therefore (hypothetically) should have allowed LD binding. Only 3 (50%) of them actually did so, but Wells interprets this as “not necessarily
problematic” (p. 243), stating that “[t]he results are consistent with [the first] hypothesis because... local binding responses are also allowed” (p. 244). Such a conclusion seems questionable, however, considering that these learners were rejecting the LD antecedents they were supposed to allow. There were 4 learners who demonstrated having acquired verbal agreement but not reflexive morphology, and who therefore (hypothetically) should have allowed LD binding across nonfinite clause boundaries but not finite ones. Only 1 (25%) actually allowed LD antecedents in nonfinite contexts; the other 3 showed exclusively local binding.

“Again,” according to Wells, using similarly questionable reasoning, “this is not necessarily a problem” (p. 243). Finally, there were 21 participants who demonstrated having acquired reflexive morphology in English, and who therefore (hypothetically) should have disallowed LD antecedents. In this group, 15 (71%) showed local binding, while 6 (29%) permitted LD binding. In this case, Wells acknowledges having found “binding preferences that were not predicted to occur” (p. 243).17

Attempting to explain the results she considered unexpected, Wells makes the interesting point that “developing systems may have characteristics that are simply not evident in mature languages” (p. 250); for instance, child L1 learners demonstrate incipient knowledge of inflectional morphology before reaching what researchers might consider to be a criterion of ‘acquisition’. She also explores the idea that the distinction between X0 and XP reflexives is not fine-grained enough and should be more than binary. For example, she argues, if learners were

---

17 Wells (1998) also carried out chi-square analyses to check for statistically significant relationships between (1) verbal morphology and binding, and (2) reflexive morphology and binding, and she interprets her results as indicating “nonaccidental” relationships in both cases (p. 244). However, she appears to have used tallies of responses as opposed to tallies of people, which violates the assumptions of this statistical test, making it unclear how trustworthy the results are. Even if the assumptions had not been violated, it would not be appropriate to imply causal relationships based on chi-squares, as Wells does when she says that “data from learners with the XP anaphor indicate an effect for agreement on binding domain” (p. 245).
to treat \textit{himself} as instantiating an incomplete set of $\varphi$-features, it might be expected to work like the similarly underspecified German reflexive \textit{sich}, which is marked for person and number, but not for gender, and which shows sensitivity to finiteness in LD binding (p. 247). Like English \textit{himself}, such reflexives do not allow modifiers (e.g., consider \textit{*stupid myself} and \textit{my stupid self} versus \textit{*dummer mich}, which has no internal place for the adjective to appear). As such, Wells reasons that some of the participants might have succeeded on her GJT assessing reflexive morphology even if their reflexives were incompletely specified. If this were true of the learners who allowed LD antecedents across nonfinite clauses while seeming to treat English reflexives as XPs, then their results might be explainable (post hoc). This is speculation, of course, and does not account for all of the results, but it is at least an interesting way of attempting to make an overly simple binary distinction finer-grained.

Another study, conducted by White (1995b), produced results similar to those in Wells’ experiment in the sense that learners of L2 English were found to allow LD binding despite showing quite accurate recognition of gender and number agreement on English reflexives. This research will be discussed below in the section on attempts at instructional intervention.

In a study assuming movement of reflexives at LF (e.g., Cole et al., 1990; Pica, 1987), Yip and Tang (1998) took an innovative approach to investigating the binding behavior of L1-Cantonese learners of L2 English. To try to gain some insights into L1 influence, they tested the participants in both their L1 and their L2. One of their goals was to assess the validity of Yuan’s (1994) ‘Positive Transfer Hypothesis’—the idea that targetlike demonstrations of L2 binding characteristics might be “explained in terms of similar properties instantiated in learners’ native languages” even if the languages differed in other ways (Yip & Tang, 1998, p. 165). Importantly,
while Chinese has a long-distance monomorphemic reflexive (ziji, ‘self’ in Mandarin; zigei, ‘self’ in Cantonese), it also has a local polymorphemic reflexive (taziji, ‘himself’ in Mandarin; keoizigei, ‘himself’ in Cantonese), which Yip and Tang reasoned could serve as a basis for acquiring local reflexives in L2 English. Importantly also, however, the polymorphemic reflexives in Mandarin and Cantonese (taziji and keoizigei) have been shown to be ambiguous: Each can function either as a local polymorphemic reflexive or as a pronoun plus an emphatic reflexive morpheme, in which case LD antecedents are possible (Tang, 1985). The upshot is that whereas English himself allows only local subject and non-subject antecedents, and whereas Cantonese zigei allows only subject local and LD antecedents, Cantonese keoizigei can allow local subject, local non-subject, and LD subject antecedents, disallowing only LD non-subject antecedents.

The fact that keoizigei shows less restricted binding behavior than English himself might seem to argue against the idea that L1-Cantonese learners of English would be able to use their knowledge of keoizigei as a means of acquiring the restrictions on English himself; nonetheless, this is the line of reasoning that Yip and Tang pursue. Hypothesizing that “those learners who have acquired the local-binding characteristics of English will also treat taziji and keoizigei as local anaphors” (p. 172)—i.e., that people showing a restriction in their L2 will also show that restriction (unnecessarily) in their L1—they ask, “do learners transfer the LD binding properties of Cantonese zigei (self), and can the properties of the local compound reflexive keoizigei serve as a basis for the acquisition of the target properties?” (p. 166). If the reasoning seems inverted, it is not that Yip and Tang are ignoring keoizigei’s long-distance possibilities, but rather that

---

18 Whereas *John+self is not possible in English, John+zigei ['John-self'] is possible in Cantonese. Yip and Tang (1998) argue that when zigei is used in this way to reinforce a name or pronoun, the compound form has a different structure from that commonly assumed for XP reflexives.
they apparently think that different L1 speakers might treat \textit{keoizigei} differently and that the way they treat \textit{keoizigei} in Cantonese should match the way they treat \textit{himself} in English.

The participants in Yip and Tang’s (1998) experiment were 268 L1-Cantonese learners of L2 English, ages 13-21, with at least 8 years of formal English instruction and exposure to the language starting around age 6. They were divided into 3 proficiency levels based on the results of a cloze test. There were also 30 native speakers of English included as controls. Yip and Tang employed 3 sentence types: biclausal with finite embedded clauses, biclausal with nonfinite embedded clauses, and monoclausal dative constructions. For each sentence type, 4 tokens contained reflexives and 4 contained pronouns. A Cantonese version of the test included sentences with \textit{zigei} and also sentences with \textit{keoizigei}. A crucial element of the scoring was that “only those learners who evidenced a fair amount of systematicity... were included” in the analyses (p. 177); that is, participants had to show the same pattern of responses on 3 out of 4 tokens of each type in order for their data to count, so to speak. The results are presented in terms of the percentages of learners at each proficiency level who answered in a certain way.

Looking at the results for L2 English, Yip and Tang found that more proficient learners were more likely to (correctly) select only local antecedents (20% of low-proficiency learners, 59% of mid, and 75% of high, versus 93% of native speakers) and less likely to (wrongly) allow LD antecedents (finite: 46% of low, 17% of mid, 15% of high, and 0% of NSs; nonfinite: 44% of low, 21% of mid, 18% of high, and 0% of NSs).\textsuperscript{19} They interpret the lower-proficiency learners’ behavior as evidence of transfer based on L1 \textit{zigei} (p. 186), construing their rate of LD binding

\textsuperscript{19} Incidentally, Yip and Tang report having found only “negligible” evidence for a finite/nonfinite asymmetry; there were only 3 low-, 1 mid-, and 2 high-proficiency learners who consistently allowed LD binding across nonfinite clause boundaries but not across finite ones (p. 180).
as being “compatible with the view that learners initially treat English anaphors on a par with the [L1] simplex anaphors” (p. 188). They also note that although most of the high-proficiency participants seemed to have acquired local binding in English, “a certain percentage of learners at each level retained LD binding and did not succeed in unlearning it” (p. 184).

Based on the results of the measure assessing participants’ interpretations of zigei and keoizigei in their L1, however, Yip and Tang do not consider the Positive Transfer Hypothesis to be clearly supported; only about half of the participants consistently required only local binding for keoizigei, and their patterns of interpreting keoizigei in L1 and himself in L2 did not necessarily match. Looking only at the learners who had demonstrated having acquired local binding in English, Yip and Tang found that “few [0-38% across levels] consistently bound both [L1] and L2 compound reflexives strictly locally.... In fact, many [34-58%] consistently allowed LD antecedents for [keoizigei], but observed local binding in English” (p. 186). Thus, it did not seem that these learners had acquired local binding “simply by reanalyzing the target reflexives as having the same properties as their [L1] counterparts” (p. 186). Yip and Tang conclude, somewhat predictably considering the characteristics of keoizigei they outlined, that “keoizigei fails to provide an adequate model [for transfer... and] indeed, it may actually have predisposed learners’ grammars in a direction that is less L2 target-like by reinforcing LD binding” (p. 187). Their approach, in sum, did not produce results indicating that L2 learners can use their knowledge of morphologically complex reflexives in an L1 as a stepping stone for developing targetlike knowledge of morphologically complex reflexives in an L2.

Other researchers, approaching this issue in perhaps a more straightforward way, have also concluded that learners may be unlikely to make use of morphologically similar L1
reflexives in interpreting reflexives accurately in the L2. Hirakawa (1990), for instance, simply asked participants to provide L1-Japanese translations for two L2-English test sentences at the end of her experiment. She found that the learners translated *himself* in a variety of ways (e.g., as *zibun, zibun-zisin, kare-zisin*, proper names, etc.) and that the characteristics of the words they used as L1 translations did not necessarily correspond to the binding properties they showed in L2 English; some participants who translated (local) *himself* as (local) *kare-zisin* did not necessarily show local binding, and some participants who performed perfectly with local binding in English translated (local) *himself* as (long-distance) *zibun*. Thus, in Hirakawa’s estimation, choices of translation equivalents “do not appear to explain either the successful or the unsuccessful learners” (p. 78, note 8). Akiyama (2002) identifies several problems with the idea that the properties of morphologically complex L1 reflexives might be transferred directly to the L2. Among them, regarding an L1-L2 pairing of Japanese and English, he mentions that “it is doubtful whether Japanese actually has a reflexive that corresponds exactly to the properties of English reflexives” (p. 46). He also points out that *kare-zisin* is not often used in conversation; rather, it is “(unnaturally) introduced to Japanese learners when they learn English *himself* in classes” (p. 46). In another article, arguing against the idea that L1-Chinese learners of L2 English might associate polymorphemic *taziji* with English *himself*, L. Jiang (2009) reasons that *taziji* has a “weaker influence on Chinese speakers” than *ziji* does since *ziji*’s antecedents can be of any person or number, and it can therefore often replace *taziji*, which is limited to third-person singular antecedents (p. 484).

In short, there does not appear to be much evidence for direct interlingual identifications between particular reflexives in the L1 and corresponding reflexives in the L2.
Whether hypothesizing about L1 influence or about purportedly universal characteristics of language, researchers have also not tended to find clear relationships between L2 binding behavior and verbal and reflexive morphology (if they have looked). Moreover, as we have seen, there are many attested exceptions to the generalization that the types of antecedents allowed for reflexives (local or long-distance, subjects or non-subjects) are tied to the reflexives’ morphological characteristics (mono- or polymorphemic). Despite this, it seems that some UG researchers are continuing to explore how the concepts of morphological triggers or overt morphological prerequisites for L2 binding behavior might be viable. For instance, while acknowledging that “this form-meaning relationship is only a tendency and not without exceptions,” Slabakova (2008) explains what is assumed in many studies:

The... visible morphological form of the anaphor can serve as a clue to the possible binding domain, and, associated with that, its orientation.... What learners [of L2 English] have to acquire... is how the English reflexives *himself*, *herself* should be classified: as monomorphemic (X₀) or as bimorphemic (XP). The relevant interpretations will follow from the universal tendency described above (pp. 129-30).

Indeed, as indicated in passing above, L. Jiang (2009) speculates as follows in seeking to explain the results of his empirical study:

Chinese speakers appear initially to misanalyse English polymorphemic reflexives as monomorphemes, due to transfer of L1 knowledge of the monomorphemic reflexive *ziji*.... At the same time, they may also transfer the properties of tense... unaware of the fact that, unlike Chinese, tense is realized morphologically in
English finite clauses.... As the level of proficiency increased, the Chinese
speakers found that tense is realized morphologically in English finite clauses,
but meanwhile failed to reanalyse English reflexives as morphologically
complex.... By advanced proficiency, the Chinese speakers realize that English
reflexives are polymorphemic and must be adjoined to the immediate VP (pp.
483-5).

This reasoning is very reminiscent of statements made by researchers a decade earlier, just two
eamples of which are given below, from Yuan (1998) and MacLaughlin (1998):

[English-speaking learners of Chinese] have to become aware of the morphological
difference between *ziji* and English reflexives.... These learners have acquired the
Chinese reflexive *ziji* as a X₀ reflexive and know that the reflexive and its binder must
have the same X-bar status.... However, these learners may transfer the properties
of AGR in their L1... unaware of the fact that, unlike English, AGR in Chinese finite
clauses is morphologically null (Yuan, 1998, p. 335).²⁰

[If learners of English] are able to... change their representation of English reflexives
to a morphologically complex one, then local binding should ‘come for free,’
assuming the learners are constrained by UG.... Presumably, [AGR] parameter
resetting could be triggered by the recognition of overt agreement morphology
found on verbs in English (MacLaughlin, 1998, pp. 217-8).

---
²⁰ To explain how it might be that L1-English learners of Chinese could fail to take AGR in Chinese as
morphologically null, Yuan (1998) suggests that they might misinterpret Chinese aspect markers, such as perfective
*le* and experiential *guo*, as “evidence for the morphological presence of Tense” (p. 335, note 4).
If it were, in fact, reasonable to assume that there might be overt morphological triggers of L2 binding behavior, then that would be important to acknowledge in the context of the present study, one of whose ultimate goals is to test S. E. Carroll’s (2001, 2007) awareness constraint on negative evidence. It will be recalled that Carroll predicts that feedback and metalinguistic information cannot function as negative evidence if surface manifestations of the relevant linguistic phenomena are not available to be processed consciously. If learners were able to acquire the syntactic and semantic properties of reflexives automatically through their recognition of overt morphological characteristics, then *zibun’s* lack of person, gender, and number features might be argued to play a role in the working of feedback independently of the fact that important structural properties (i.e., subjecthood, c-command) are not in one-to-one correspondence with surface characteristics, such as nominative case-marking or sentence-initial position.

*Morphological triggers?*

A concern that seems to arise repeatedly for UG researchers attracted to the notion that (awareness of) L2 morphology can drive L2 syntax is expressed by L. Jiang (2009), who notes that “it might seem odd that [L2] input providing transparent evidence showing that reflexives are morphologically complex fails to trigger the relevant reflexive type” (p. 285). There are, however, several strands of research whose findings might be taken as shedding light on this and related issues. They include (but are not limited to) studies suggesting that L2 learners may not initially process or use overt morphology like native speakers do even when the L1 and L2 systems work similarly on some level (e.g., Foucart & Frenck-Mestre, 2010; Slabakova & Gajdos,
(2008), and studies indicating that learners with many years of L2 exposure may demonstrate high levels of metalinguistic knowledge of morphology, as well as apparently nativelike uses of morphology in certain situations, without showing sensitivity to morphological markings in all situations of real-time sentence processing (e.g., on self-paced reading tasks checking for sensitivity to plural and/or agreement markings [N. Jiang, 2004, 2007], or in eye-tracking experiments involving gender-agreement violations at various structural distances [Keating, 2009]; see also Clahsen & Felser, 2006b). It has also been reported that a learner with many years of L2 exposure might demonstrate quite inaccurate use (or non-use) of overt morphology (appearing in that regard to have stabilized far short of targetlike accuracy), and yet perform very accurately with aspects of syntax that are theoretically related to those morphological features (e.g., Lardiere, 1998b).

In fact, an active and growing area of theoretically driven research is currently seeking to explain why L2 morphological development so often falls short of targetlike. One debate revolves around the question of whether L2 learners might suffer “representational deficits” of certain types of features which are not available from the L1 (according to the Failed Functional Features Hypothesis [e.g., Franceschina, 2001; Hawkins & Chan, 1997] and the Interpretability Hypothesis [Tsimpli & Dimitrakopoulou, 2007; Tsimpli & Mastropavlou, 2007]) or whether learners might have the features represented at an abstract level but simply fail to realize them overtly (the Missing Surface Inflection Hypothesis [e.g., Hazdenar & Schwartz, 1997; Prévost & White, 2000; White, 2009; see also Lardiere, 1998b, 2008, 2009]). Another overlapping debate focuses on issues of online sentence comprehension and the proposal that L2 learners might rely predominantly on lexical, semantic, and pragmatic information when processing sentences...
in real time, differing from native speakers in that they do not compute full syntactic parses. According to Keating (2009), whereas Clahsen and Felser (2006a, 2006b), in their Shallow Structure Hypothesis, would attribute failures in parsing to grammatical deficits (assuming that “the grammar that feeds the parser diverges in some way from the native grammar”), other UG- and processing-/processability-oriented researchers (e.g., Gregg, 2001; Pienemann, 1998, 2005; Sorace, 2006; Truscott & Sharwood Smith, 2004; VanPatten, 1996, 2004) assume that learners lack “appropriate parsing heuristics” for the L2 (p. 507).

A full review of this extensive literature is beyond the scope of this dissertation. Suffice it to say that even when morphology seems, on some level, to be “transparent,” that does not mean that L2 learners will internalize it in such a way that it functions as it does for native speakers. A morpheme might mean something metalinguistically, while functionally it amounts to noise. From N. Jiang’s (2007) perspective, some types of L2 knowledge might become “an integral part of the learner’s mental representation that is automatic in its activation and functioning” (p. 2), but other types (such as, perhaps, certain types of inflectional morphology) might “stabilize in a non-native state” and never be completely integrated (p. 19; see also Long, 2003). Slabakova (2009) points out that “even if there is a complete match of features, morphological acquisition is by no means a walk in the park.... [E]ven simple mapping between L1 and L2 morphemes carrying exactly the same grammatical meanings can... present considerable difficulties” (p. 322). And matching features would seem to be a best-case scenario. Lardiere’s (2008, 2009) Feature Reassembly Hypothesis highlights the complexities and difficulties L2 learners face in identifying and (re)configuring or (re)mapping features—figuring out which functional categories they might be associated with in the L2’s syntax and
how they combine into feature matrices associated with lexical items which might be qualitatively from those in the L1.

According to Lardiere (2009), the reconfiguration of features for an L2 is “a formidable learning task that goes far beyond the simple ‘switch-setting’ or ‘selecting’ metaphors” of parameter resetting (p. 175). For one thing, specifications of feature values may need to be relativized: If an L2 has morphological forms distinguishing among singular, dual, and plural, for instance, its [+plural] feature might not be comparable to a [+plural] feature in an L1 with morphological forms distinguishing only between singular and plural. This insight seems applicable across many areas of SLA, including for whichever morphological features are hypothesized to be relevant to the binding behavior of reflexives. Morphological characteristics that seem to be analogous across languages in one sense might play different (relativized) roles in the context of differing linguistic systems (L1s versus L2s) as a whole, and they might play different (relativized) roles within the particular sub-areas of language which linguists have attempted to carve out for focused analyses (e.g., of reflexive binding).

Slabakova (2009) admits, “We all knew that the rosy view of parameters being responsible for a range of superficially unrelated constructions appearing in the interlanguage grammar at the same time was too good to be true anyway” (pp. 313-4). Nevertheless, she warns against drifting toward an “assumption that the grammar of a language is an assembly of different grammatical constructions with no inherent similarities computable by parameters” (p. 315). She emphasizes that “language variation is not limitless” and that it is precisely the predictable constraints on L2 behavior that make SLA interesting as a cognitive science (p. 323). Predictable constraints are, indeed, quite fascinating when they can be demonstrated. However,
as far as the idea of morphological triggers in adult SLA is concerned, there appear to be reasons to believe that abstract morphological features, overt morphological and syntactic production, and online morphosyntactic processing do not always coincide. This being the case, while it seems very reasonable for researchers focusing on L2 learners’ knowledge of reflexives to test their knowledge of morphology directly as opposed to speculating about it, it does not seem very reasonable to assume that this morphological knowledge will be associated with binding characteristics in the same way as is theorized for the L1.

In this regard, it may be worth asking exactly what Yuan (1998) and L. Jiang (2009) mean when they propose that learners have to become “aware” of morphological differences between reflexives across languages; what MacLaughlin (1998) means when she says that parameter resetting could be triggered by “recognition” of overt agreement morphology; and why L. Jiang (2009) considers it possibly odd that “transparent evidence” of morphology does not “trigger” theoretically related characteristics. If UG triggers were thought to work instantaneously via learners’ identifications of overt morphology (metalinguistic or not), then a finding that even quite “transparent” morphology did not function as a “trigger” might indeed seem odd. At least one problem here seems to be that the terminology and metaphors we use can influence our reasoning processes. The metaphor of a trigger connotes instantaneity; words like recognize and realize are punctual achievement verbs; words like misanalyze and reanalyze connote conscious top-down processes. It seems highly unlikely that UG researchers would intend to imply that learners’ metalinguistic knowledge can immediately guide the way their linguistic competence functions. There nonetheless seems to be a danger of unintentionally internalizing counter-theoretical aspects of words’ meanings and falling into a trap of literal
interpretation even if, upon reflection, one would disagree with the implications. In reality, even beginning learners are likely to “recognize” and be “aware” of reflexive and verbal morphology if they have received classroom instruction. Although teachers do not tend to refer explicitly to theoretical constructs such as AGR and VP adjunction, many do certainly draw attention to person, number, and gender morphemes. Conversely, even advanced learners are unlikely (unless they are linguists) ever to “understand” that the X-bar status of a reflexive and its antecedent must match or to “realize” that English reflexives must be adjoined to an immediate VP. This is not just splitting hairs; if metaphors or imprecise shorthand influence our thinking in ways that gainsay our intended conceptions of theoretical constructs, that can lead to apparent contradictions as well as to faulty conclusions and specious new hypotheses.

Attempts at instructional intervention

Considering the mixed empirical results regarding the question of L1 influence, it is not actually clear that learners of a given L2 will always need the specific kinds of positive or negative evidence that one might predict based on translation equivalents in the L1. However, it does definitely seem to be the case that L2 learners do not interpret sentences with reflexives accurately right away and that they persist in making errors even at advanced levels of proficiency. Regardless of how precisely it is possible to predict the binding patterns learners will actually show, and regardless of how credibly current linguistic theories can explain the reasons for them, the fact remains that reflexive binding is quite a problematic area of acquisition for adult L2 learners. Assessing the state of affairs from their point of view, Slabakova (2008) points out:
We ought to acknowledge the tremendous complexity of the learning task and the insufficient evidence learners have to work with. By insufficient evidence, I mean not only the general lack of consistent instruction and negative evidence, but more importantly, the general permissiveness of UG-sanctioned options, if we include misclassifications of reflexives as pronouns and logophors. This being the case, it is not surprising at all that learners have a very hard time acquiring and consistently interpreting reflexives in a second language, as the existing research studies have amply documented. Reflexive binding is an area at the syntax-semantics interface where too many available options are provided by UG (pp. 133-4).

From Slabakova’s perspective, in this area of SLA, even if UG might play a role, L2 learners are “armed only with a (non-exceptionless) universal tendency” (p. 133). A natural question in such a scenario, consequently, is whether the process of learning can be made more efficient through instruction.

Two studies have attempted to take advantage of insights from linguistic theory in order to facilitate the acquisition of reflexives: an experiment with learners of L2 English by White (1995b) and an experiment with learners of L2 Japanese by White, Hirakawa, and Kawasaki (1996). Both explored whether providing positive evidence of one characteristic would amount to a two-for-one deal, indirectly activating a theoretically related characteristic for free, so to speak. In the former case, the idea was that positive evidence of object antecedents for English reflexives might indirectly serve as negative evidence against long-distance antecedents. In the latter case, it was that positive evidence of long-distance subject antecedents for Japanese
zibun might indirectly serve as negative evidence against long-distance object antecedents (or at least not induce learners to overgeneralize that they were allowable).

Assuming that, due to transfer, L1-Japanese learners of English might be expected to show LD reflexive binding in the L2 with an orientation to subject antecedents, White (1995b) asked, “What properties of English L2 input could reveal... that such assumptions are incorrect?” (p. 67). Although positive evidence of non-subject antecedents might be available, she argued that LD binding posed a potential problem: Evidence for local antecedents would “not be sufficient to show that long-distance binding is ruled out” (p. 67).²¹ In order to explore a possible answer to her question, she conducted an experimental study with 30 adults studying English at an intermediate level, 11 of whom spoke Japanese as their L1 and 19 of whom were native speakers of French. White also included a control group of 20 native speakers of English. In addition to her hypothesis that the Japanese speakers would initially allow LD antecedents for reflexives but restrict them to subjects, White predicted that “[p]roviding positive evidence that English in fact allows binding to objects (either explicitly or implicitly) should lead to knowledge not just that binding to objects is possible but also that long-distance binding is impossible” (p. 67).²² Now, as we have seen, there are languages that allow local binding to objects along with LD binding to subjects, as well as languages that allow LD binding with

²¹ Many other researchers have also reasoned this way. For example, according to Akiyama (2002): “There should be no problem when L2 learners discover that English reflexives allow local binding because they are constantly exposed to English sentences that show that this is the case... by means of positive evidence. However, matters would not be so simple when [L1] Japanese learners acquire... that LD binding is impossible.... [T]here seems to be no evidence available to them indicating that this restriction exists in English.... [and] the L1 (Japanese)... simply has no such restriction.... Furthermore... it is quite rare that learners are explicitly told or taught about such a restriction” (pp. 30-1). Likewise, according to L. Jiang (2009): “English input may provide ample evidence showing that local binding is possible, but this by no means suggests that LD binding is impossible in English” (p. 470).
²² Incidentally, White (1995b) also suggests that “[a]nother kind of positive evidence that might serve the same purpose is the morphological form of the reflexive itself. If English reflexives are analyzed by L2 learners as morphologically complex... they should realize that an XP anaphor is involved and hence that long-distance binding is prohibited” (p. 67).
morphologically complex reflexives, so in that sense, White’s study may have been doomed from the outset. Nonetheless, it is valuable to review the methods she employed and the results she found.

White (1995b) divided her participants into 3 treatment groups. One received explicit instruction (referred to by White as “explicit positive data”) over a period of 4 weeks, with three 20-minute sessions per week focusing on the grammar of reflexives (p. 68). This instruction informed the learners about the morphological forms of reflexives in English (i.e., marking of gender and number) as well as about the fact that English reflexives can take non-subject antecedents. However, no mention was made of the fact that LD antecedents were not allowed. Another set of learners received “naturalistic positive data” (p. 69); likewise over a 4-week period with three 20-minute sessions per week, they answered comprehension questions based on reading passages which included examples of reflexives bound both to subjects and to objects. The third group was not given any instruction or reading passages involving reflexives. White’s assessment measures included a self-paced GJT to check whether the learners knew reflexives’ gender- and number-agreement properties, and a story-based truth-value judgment task with 8 items for each of 5 sentence types: monoclausal items testing for subject antecedents, monoclausal items testing for object antecedents, biclausal items with finite embedded clauses, biclausal items with nonfinite embedded clauses, and monoclausal items testing for strategies based on linear order, such as choosing the closest antecedent.

In analyzing the results of the pre- and post-tests, White (1995b) found no statistically significant effects for L1, treatment type, or the interaction between L1 and treatment type, and, in her words, “no indication that any treatment led to improvement” (p. 73). The small
sample size of the study must be kept in mind, but even with more participants the magnitude of the effect may not have been very meaningful: The average amount of change in either direction, on a test with 40 items, ranged from 0.5 to 2 points, except for an improvement of 5 points by an L1-Japanese speaker in the no-treatment condition (the only Japanese speaker in that condition). White concludes, “On the face of it, the results are rather discouraging, since neither of the initial hypotheses was supported” (p. 74); “[t]here was certainly no indication that positive evidence on binding to objects triggered knowledge of the impossibility of long-distance binding” (p. 75). However, White rightly points out that the SLA research community’s understanding of L2 knowledge and learning can benefit just as much from negative results as from positive ones (p. 77). In addition to finding no differences between the different L1 groups (e.g., suggesting that transfer does not seem to have been involved), White also did not find evidence of a subject orientation on the pre-tests, suggesting that the participants might not have needed positive evidence of binding to objects in English in the first place. Only two learners (1 Japanese and 1 French) showed a pattern of rejecting objects and allowing LD antecedents; many allowed both objects and LD antecedents both before and after the treatments. Moreover, the results of the GJT indicated that the participants were sensitive to reflexive morphology; they were “quite accurate on gender and number agreement” (p. 75), calling into question the adequacy of the theory of reflexive binding used as a foundation for the experiment. White admits that the idea of instantaneous parameter setting is “clearly an idealization even in L1 acquisition” (p. 76).

Undaunted, White et al. (1996) conducted another experiment to investigate “whether specific teaching on one property of the Japanese reflexive *zibun*, namely the fact that it can
take long-distance antecedents, leads to knowledge of another property, namely that LD antecedents must be subjects” (p. 236). They hypothesized that acceptance of LD subject antecedents could be taught via positive evidence and that it would indeed be accompanied by a rejection of LD object antecedents; that is, they did not expect the learners to overgeneralize to an ‘impossible’ grammar in which LD objects were allowed. The researchers recognized that there are languages whose reflexives allow both local objects and LD subjects as antecedents and, as such, did not hypothesize that teaching about LD antecedents in Japanese would lead the learners to reject local objects. Importantly, White et al. (1996) also argued against the idea that zibun’s morphological properties might serve as a trigger for other linguistic characteristics in this learning situation, stating in a footnote that, considering the exceptions that exist cross-linguistically, “there is no clear implicational relationship in the relevant direction between the head property and the LD property... or between the head property and subject orientation” (pp. 238-9, note 5). Since, under this theoretical framework, all learners, regardless of L1 background, would presumably already ‘know’ that LD antecedents cannot be objects, it seems imprecise to have asked whether positive evidence of LD subject antecedents could ‘lead to’ this knowledge; however, it was still worth asking whether overgeneralizations might occur.

The participants in White et al.’s study were 13 learners of L2 Japanese from a variety of L1 backgrounds (5 Chinese, 4 French, 2 English, and 2 Korean), studying the language in a low-intermediate university course. (Half of them spoke L1s that, like Japanese, allowed LD binding to subjects, but this turned out apparently not to be an issue.) Also included were 10 native speakers of Japanese as controls. One Chinese-speaking L2 learner and one Japanese native speaker were eliminated due to apparent response biases and poor performance on distractor
items, leaving 12 L2 learners and 9 native speakers as participants. The instructional treatment took place over a period of 4 weeks (about 15-20 minutes per day, 2 or 3 times per week, for a total of about 3 hours) and involved exposure to positive evidence through reading and writing exercises, sometimes accompanied by pictures. The learners were told explicitly that sentences with *zibun* could sometimes be ambiguous and that antecedents could appear outside the clause in which *zibun* appeared; however, the teaching materials never used object antecedents, and no mention was made of *zibun*’s subject orientation. White et al. administered a pre-test and 2 post-tests: one immediately following the teaching intervention, and another about 5 weeks later. These were in the form of picture-based truth-value judgment tasks with 4 items for each of 5 sentence types: monoclausal sentences testing subject or object antecedents, and biclausal sentences testing local subject, LD subject, or LD object antecedents. There were also 18 distractors (monoclausal and biclausal sentences without reflexives) for a total of 38 items.

Group-based analyses revealed that, like the native speakers, the learners of L2 Japanese correctly accepted local subjects as antecedents in monoclausal sentences and never tended to accept objects as antecedents, whether local or long-distance, even on the pre-test. Unlike the native speakers, the L2 learners tended initially to reject LD binding to subjects, but they began to accept such antecedents following the treatment. Individual-based analyses revealed, similarly, that most of the participants responded consistently and showed high levels

---

23 Somewhat unexpectedly, in biclausal sentences, the learners were more likely than the native speakers to accept (allowable) local subjects as antecedents; that is, they appeared to answer in a more ‘nativelike’ fashion (theoretically speaking) than the NSs did. The mean acceptance response rate for NSs was only 1.89/4, with 4 NSs consistently accepting local subject antecedents and 5 NSs consistently rejecting them. As will be discussed in Chapter 4 in relation to piloting the experimental materials with NSs of Japanese, it has frequently been found that Japanese speakers may fail to recognize local subjects as antecedents for *zibun* when LD antecedents are also available, even though the grammar allows both (e.g., Akiyama, 2002; Hirakawa, 1990; Thomas, 1991).
of accuracy regarding the acceptability of local subject antecedents (in monoclausal and biclausal sentences) and the unacceptability of object antecedents (local and long-distance). For the items involving LD subject antecedents, the learners tended to show patterns of consistently accepting them (correctly) or rejecting them (incorrectly) rather than inconsistent, “wavering” responses (p. 245). On the pre-test, 1 learner consistently accepted LD antecedents, 8 consistently rejected them, and 3 were inconsistent. On the first post-test after the treatment, there was improvement: 7 consistently accepted LD antecedents and 4 consistently rejected them. The delayed post-test showed a similar pattern.

White et al. consider it noteworthy that none of the participants appeared to operate with an English-like grammar which allowed both local subject and object antecedents; instead, the predominant grammar type at the outset of the study was one that allowed only local subjects. (The authors acknowledge, though, that several studies have found a tendency for NSs not to interpret local objects as antecedents for reflexives in English despite the fact that they are allowed [e.g., Hirakawa, 1990; White et al., 1997].) While advising caution in interpreting their findings due to the small sample size, White et al. conclude that “the attempt to trigger various UG properties via L2 classroom input was partially successful” since half (7) of the participants acquired LD binding and nearly all of them refrained from overgeneralizing this to LD object antecedents (p. 251). Regarding the question of transfer, the researchers report no identifiable differences related to L1; most participants assumed that *zibun* allowed only local subject antecedents on the pre-test, and the learners who demonstrated consistent acceptance of LD antecedents on the post-tests were fairly evenly split across the 4 L1 groups.
The results of these two studies are not definitive judgments, of course, on the effectiveness of instruction on reflexives in general. White and her colleagues investigated whether targeting one aspect of an L2 system of reflexives might bring about more accurate performance in an untargeted area, but it is also interesting to ask whether learners can benefit from feedback that contains metalinguistic information regarding how the L2 system of reflexives works as a whole. In exploring this, it will be important to keep in mind an insight from Bruhn-Garavito (1995): “If ‘natural’ input underdetermines the complex target grammar, explicit input does so even more. Our understanding of language is still very tentative” (p. 80). Particularly with L2 reflexives as a target, it is crucial to recognize that linguistic theories of binding “leave certain swaths of facts unexplained” (Slabakova, 2008, p. 133). Still, as White, Bruhn-Garavito, Kawasaki, Pater, and Prévost (1997) point out, “Regardless of which theory one adopts to explain the interpretation of reflexives, certain sentences involving reflexive pronouns are ambiguous” (p. 147). At the very least, researchers can attempt to determine which readings of particular sentence types native speakers tend to accept and reject, then target those types and observe how much L2 learners tend to improve (or not) when given particular kinds or amounts of information.
CHAPTER 4

METHODS

Research questions

CALL research into the effects of more versus less explicit feedback types has produced findings in need of further investigation and explanation: In some studies, learners in explicit treatment conditions who are provided with metalinguistic information concerning the linguistic target(s) outperform those in less explicit conditions who are simply told whether their answers are right or wrong (e.g., Rosa & Leow, 2004b), whereas in other studies, no such advantage is found (e.g., Sanz & Morgan-Short, 2004), or the more explicit groups are observed not to maintain their gains (Bowles, 2005; Lado, 2008). A variety of linguistic targets have been tested, and a variety of feedback types have been employed; however, no study to date has attempted to use visual diagrams as a way of illustrating structural characteristics of sentences (e.g., c-command, subjecthood) which are relevant to producing targetlike interpretations.

Meanwhile, increasing numbers of studies are indicating that L2 learners with different profiles of characteristics (e.g., in grammatical sensitivity, working memory, motivation, etc.) respond differently to different types of L2 instructional conditions. What researchers might initially interpret as an overall lack of a treatment effect can amount, in actuality, to evidence that what works for certain learners does not work for others (e.g., DeKeyser, 1993). This makes it all the more important—especially, perhaps, in utilizing a novel and highly abstract form of feedback—to measure a variety of individual differences and to analyze their relation to learning success in treatments with features that seem likely to draw on those abilities.
With these concerns in mind, the present investigation was guided by the following research questions:

1. **Feedback**: Do learners show improved accuracy in their interpretations of L2 sentences containing reflexives after engaging in one of the following types of practice:
   
a. without feedback on the accuracy of their interpretations,
   
b. with feedback indicating the accuracy of each interpretation, or
   
c. with feedback indicating the accuracy of each interpretation and providing a tree diagram illustrating the underlying structure of the sentence?

   If so, is one or more of these treatment conditions more effective than another?

   (Does feedback containing tree diagrams help?)

2. **Individual differences**: Are individual differences in L2 learners’ grammatical sensitivity, rote memory for linguistic material, visual short-term memory, sensitivity to ambiguity in English, knowledge of metalinguistic terminology, enjoyment of grammar, and previous study of Japanese and linguistics related to their accuracy in interpreting L2 sentences containing reflexives?

3. **Aptitude-treatment interactions**: Do the relationships among these individual differences and learners’ accuracy in interpreting L2 sentences containing reflexives differ according to treatment condition? If so, which individual differences are the most strongly related to performance in each treatment condition?
Participants

The participants in this study were English-speaking learners of Japanese as a foreign language (JFL), recruited from university Japanese courses at the intermediate level or above. Of the 138 students who initially signed up to participate, 113 showed up to the first research session, and 84 completed enough activities for their data to be included. (In other words, 29 discontinued their participation at some point.) The data of 4 participants were discarded due to early-childhood experience with Japanese and/or high pre-test scores, as discussed below in the section on test scoring. This left 80 participants whose data were included in most, if not all, of the analyses. In the few cases where a participant had failed to complete a particular measure, s/he was excluded only for those analyses where that measure was relevant.

The 80 participants ranged in age from 17 to 36 years old, with a mean age of 20.94 (standard deviation: 2.83). There were 42 female participants and 38 male. On a background questionnaire administered on the first day of the experiment, 60 of the participants reported only English as their native language; 9 reported speaking English plus one other language natively (3 Spanish, 2 Bengali, 1 Cantonese, 1 French, 1 Tagalog, and 1 Tamil), and 2 reported speaking English, Spanish, and one other language natively (German or Arabic). The remaining 9 participants reported speaking a language other than English natively: 2 Korean, 2 Chinese (unspecified), 1 Mandarin, 1 Arabic, 1 Hebrew, 1 Gujarati, and 1 Spanish. Considering previous findings (e.g., Thomas, 1995, White et al., 1996) that participants with early exposure to long-distance anaphora did not show an advantage with the long-distance properties of *zibun* in Japanese, no participants have been excluded at this point solely on the basis of native
language (unless that language was Japanese); instead, pre-test performance was used to determine who should be excluded. However, the issue could be pursued in future research.

All of the participants were studying Japanese at universities in or near Washington, DC: 39 were taking courses at the University of Maryland at College Park, 21 at Georgetown University, 17 at The George Washington University, and 3 at The American University. Due to practical constraints, the learners’ Japanese language proficiency was not measured directly using an independent test. However, information on their Japanese course levels and length of Japanese study was gathered. Regarding course levels, 20 were enrolled in the second semester of their university’s second-year Japanese language courses (or, in 2 cases, in the summer equivalent of this), 19 were in the first semester of their university’s third-year courses, and 21 were in the second semester of their university’s third-year courses. In the latter group, 4 were simultaneously taking a Kanji and Composition course, and 2 were taking Introductory Japanese Linguistics. A further 19 participants had completed or tested out of their university’s first three years of Japanese language courses; 9 of these 19 were enrolled in the first semester of their university’s fourth-year courses (Integrated Advanced Japanese or Advanced Conversation and Composition, and/or Readings in Modern Japanese, Introduction to Bungo, or Business Japanese); 3 were taking the aforementioned Kanji and Composition course (for which the second-semester third-year course was listed as a prerequisite) plus at least one other

24 In Thomas’ (1995) study, among the 34 low-proficiency learners of Japanese who participated, 74% “were either native speakers of, or had had early exposure to, languages (largely Chinese or Korean) with long-distance anaphors that have properties like that of zibun” (p. 229, note 12), and yet the higher-proficiency learners of Japanese performed much better with long-distance anaphora. In the current experiment, the average score on the pre-test for the 70 participants who spoke English natively and had not grown up speaking either Chinese or Korean was 58.78%, with a standard deviation of 10.94 and a range from 32.14 to 78.57%. For the 6 participants who mentioned either Chinese or Korean as (one of) their native language(s), the average pre-test score appeared to be somewhat higher, at 66.75%, with a standard deviation of 11.02 and a range from 50.00 to 76.67%.
advanced Japanese course (e.g., Business Japanese, Japanese Literature in Translation, Introductory Japanese Linguistics), and were thus counted, for the purposes of data analysis, as being in an equivalent to the first semester of fourth-year courses; 1 was in the second semester of the university’s fourth-year courses; and 6 were taking a variety of advanced Japanese courses offered during that second semester (e.g., a Senior Seminar, Understanding Japan through Manga, The Art of Translation, Japan from the Margins, History of the Japanese Language). One final student had not taken Japanese courses for a while and was enrolled only in a Business Japanese course normally taken by students at or past the university’s third year of courses. In brief, the participants represented a fairly wide range of Japanese course levels, intermediate through advanced, as summarized in Table 4.1.

Table 4.1. Number of participants at each course level

<table>
<thead>
<tr>
<th>Course level</th>
<th>Year 2, Sem 2</th>
<th>Year 3, Sem 1</th>
<th>Year 3, Sem 2</th>
<th>Year 4, Sem 1</th>
<th>Year 4, Sem 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of students</td>
<td>20</td>
<td>19</td>
<td>22</td>
<td>12</td>
<td>7</td>
</tr>
</tbody>
</table>

When asked to estimate the number of years they had been studying Japanese (via a background questionnaire; see below and Appendix K), the participants’ answers ranged from a minimum of 1.5 to a maximum of 11 years, with an average of 4.3 years. Regarding the amount
of time they had spent in Japan, answers ranged from no time at all to 6 years.\textsuperscript{25} The average amount of time spent in Japan was roughly 7 months.\textsuperscript{26} Most of the participants (38) reported having spent less than a month in Japan, 22 had spent between 1 and 6 months there, 9 had spent between 6 months and a year, and 10 had spent more than a year.

The contexts in which the participants had studied Japanese were also varied: 3 mentioned elementary school, 4 middle school, 9 private tutoring and/or independent self-study, 29 high school, 4 community college, and 3 a language school in the United States (the Japan-America Society of Washington or the Washington Language Academy). Roughly a quarter of the participants (21) had studied in Japan, with 9 of them mentioning at least 1 semester at a Japanese university; 6 mentioning a specific summer program, such as Princeton in Ishikawa, intensive courses at the International Christian University in Tokyo, courses at Genki Japanese and Culture School (GenkiJACS) in Fukuoka, or courses at the Eurocenter in Kanazawa; and the rest simply mentioning an amount of time spent in Japan. Of course, all of the participants mentioned university study, and many mentioned more than one context.

The number of hours the participants reported spending on non-course-related Japanese use outside of class each week ranged from 0 to “50+”. On average, it was about 4 hours, but this number should be taken with a grain of salt; in many cases, it had to be estimated based on a statement such as “very rarely; sometimes use for about half hour in club meetings that meet every other week” (which was counted as 15 minutes per week) or “I meet

\textsuperscript{25} The participant reporting 6 years of residence in Japan had attended an English-speaking school in that country from seventh grade through high-school graduation.
\textsuperscript{26} In some cases, the numbers entered into this calculation were based on an estimate; for example, “one summer” was counted as 3 months, and an academic year was counted as 10 months, August through May. One participant stated only that he had been to Japan, but did not specify for how long; thus his response was not entered into the calculations presented here.
with a language partner once a week” (which was estimated as 1 hour). When a participant gave a range (such as 7-15), the average (11) was used in the calculations. When a participant specified a number, that number was used even if it had been qualified; for example, “at most 3” counted as 3, “at least 10” counted as 10, and “virtually none” counted as 0. Someone who reported that s/he “occasionally watch[ed] 2 or 3 hours of Japanese TV a week for listening practice” was counted as spending 2.5 hours, even though this was not always the case.27

Most participants reported multiple reasons for studying Japanese. The most commonly mentioned were a general interest in or enjoyment of languages (38); an appreciation of Japanese culture and/or the desire to learn more about it (32); a sense of curiosity, novelty, variety, uniqueness, and/or fun (19); and a desire to use the language in their future professions (18). Twelve participants mentioned wanting to live, study, work, and/or travel in Japan, while 11 mentioned family or other personal connections, and/or having lived in Japan in the past. Between 6 to 8 participants (separately) mentioned each of the following: academic requirements at their university; a sense of continuity or not wanting to lose something they had already invested time and effort into; interests related to entertainment, such as video games, music, and pop culture; scholarly or literature-related interests; a general desire to be able to communicate with people; and a desire to challenge themselves and be satisfied with achieving something they had pursued. Finally, 2 participants indicated that they had either chosen to study Japanese on a whim or mentioned that it had simply “happened to be available” during the registration period.

27 In future research, it would be advisable to provide clear ranges of numbers on a questionnaire and ask participants to circle the closest estimates.
**Linguistic target**

The linguistic target chosen for this investigation was structural restrictions on interpretations of the Japanese reflexive *zibun* (‘self’). For speakers of English learning Japanese as a second language, Slabakova (2008) describes this learning situation as “a perfect example of a mismatch at the syntax-semantics interface” (p. 129) and a “very difficult area of acquisition” at that (p. 132). In theory, a major aspect of the learning problem relates to the generalization that English *himself/herself* must be bound locally, but can take subjects or non-subjects as antecedents, whereas Japanese *zibun* can be bound long-distance (LD), but allows only subjects as antecedents (Slabakova, 2008; Thomas, 1995; White, 1995a, 1995b). What this difference means for English-speaking learners of Japanese (or for Japanese-speaking learners of English, for that matter) is that for certain kinds of sentences, interpretations of superficial translation equivalents in the two languages will be reversed.

Consider examples (1) and (2) below. The (a) version of each shows the possible interpretation(s) in Japanese, while the (b) version shows the interpretation(s) in English.

(1) a. *Yoosuke*-ga [Wataru*-ga *zibun*/*j*-o seme-ta]-to omot-ta
   
   Yosuke-NOM Wataru-NOM self-ACC blame-PAST COMP think-PAST
   
   ‘Yosuke thought that Wataru blamed self/*j’ [ambiguous in Japanese]

   b. *Yoosuke* thought that Wataru blamed himself/*j* [unambiguous in English]

(2) a. *Mikai*-ga Sakiko-ni *zibun*/*j*-no syasin-o mise-ta
   
   Mika-NOM Sakiko-DAT self-GEN picture-ACC show-PAST
   
   ‘Mikai showed Sakiko’s picture’ [unambiguous in Japanese]

   b. Mikai showed Sakiko a picture of herself/*j* [ambiguous in English]
In (1a), the Japanese sentence is ambiguous: *zibun* can refer either to the local subject, *Wataru*, or to the long-distance subject, *Yosuke*. On the other hand, (1b) in English is unambiguous: *Himself* can refer only to the local subject, *Wataru*. In (2a) and (2b), the opposite situation holds. The Japanese sentence in (2a) is the unambiguous one, admitting only the reading in which *zibun* refers to the subject, *Mika*. In the English sentence in (2b), more than one interpretation is possible: *Herself* can refer either to the subject, *Mika*, or to the indirect object, *Sakiko*.

In addition to the issues of distance and subjecthood, conditions on the interpretation of reflexives in both languages have been argued to involve the structural relationship of c-command\(^{28}\), as shown in (3) and (4) below:

(3)  
\begin{align*}
\text{a. } & [\text{Syoozoo-no sensee}]^j-ga \text{ zibun}^*_{i/j} \text{-o hihansita} \\
& \text{Shozo-GEN teacher-NOM self-ACC criticized} \\
& \text{‘[Shozo’s teacher], criticized himself}^*_{i/j}\text{’}
\end{align*}

\begin{align*}
\text{b. } & [\text{Shozo’s teacher}], \text{criticized himself}^*_{i/j} \text{ (adapted from Aikawa, 2002, p. 159)}
\end{align*}

(4)  
\begin{align*}
\text{a. } & [[\text{Chuck-ga ketta}]^j-inu] \text{-ga zibun}^*_{i/j} \text{-o kiratta} \\
& \text{Chuck-NOM kick-PAST dog-NOM self-ACC hate-PAST} \\
& \text{‘[The dog [that Chuck kicked]], hated himself}^*_{i/j}\text{’}
\end{align*}

\begin{align*}
\text{b. } & [\text{The dog [that Chuck kicked]], hated himself}^*_{i/j}
\end{align*}

In both (3a) and (3b), the reflexive can refer to *sensee/teacher*, but not to *Shozo*, since *Shozo* does not c-command the reflexive. In both (4a) and (4b), the reflexive can refer to *the dog* (that Chuck kicked), but not to *Chuck*, since *Chuck* does not c-command the reflexive. Figure 4.1

\(^{28}\) A fairly standard definition of c-command, from Chomsky & Lasnik (1993), is that “\(\alpha\) c-commands \(\beta\) if \(\alpha\) does not dominate \(\beta\) and every \(\gamma\) that dominates \(\alpha\) dominates \(\beta\)” (p. 518, cited in Hicks, 2009, p. 4).
presents a simplified diagram of (4a), showing that Chuck does not c-command zibun and therefore cannot serve as its antecedent.

Figure 4.1. C-command requirement on antecedents of reflexives

To summarize, then, zibun and himself are similar with regard to the c-command requirement, but they differ with regard to subject orientation and the allowable distance of antecedents. Japanese zibun is generally considered to be structurally more permissive than English reflexives in that it allows long-distance antecedents outside the clause in which it appears, while it is considered structurally more restrictive in that it disallows non-subjects as antecedents. English reflexives work in the opposite way: They disallow long-distance antecedents while allowing non-subjects.
All of the testing and treatment materials for the current study were designed to exemplify these syntactic characteristics and were pilot-tested with native speakers of Japanese to verify that any allowable antecedents for zibun within each sentence were, in fact, c-commanding subjects. This focus on syntactic considerations followed the tradition of a long line of generative SLA research (as discussed in Chapter 3). However, it would be remiss not to mention that it is actually controversial whether zibun is strictly subject-oriented in all contexts. Counter-examples have led some researchers to argue that “the nature of zibun-binding cannot be fully explained in terms of syntax, and pragmatic factors must be taken into consideration” (Aikawa, 2002, p. 161; see also, e.g., Iida, 1996; Kuno, 1973, 1978; Kuroda, 1973; Sells, 1987). For example, there are situations in which reflexives can be used emphatically or logophorically, in which cases they are subject to different distributional requirements (Reinhart & Reuland, 1991). None of these pragmatic factors were focused on in the present study, but it is important to be aware of their existence. More details on the similarities and differences between the English and Japanese systems of reflexives are provided in Appendix F3.

Restricting their focus to syntactic generalizations, Slabakova (2008) and many other researchers (e.g., Akiyama, 2002; Thomas, 1995; White et al., 1996) have argued that an L1-L2 pairing of English and Japanese amounts to a poverty-of-the-stimulus learning situation where reflexives are concerned. Certain “native interpretations” (allowed in English, in this case) are not allowed in the L2, so there is no positive evidence for them in the L2 input. For instance, a native speaker of Japanese would never use the sentence Mika-\-ga Sakiko\-ni zibun\-\-no syasin\-\-o mise-ta to refer to Sakiko’s picture. However, the argument goes, it is not certain that L2 learners of Japanese will take this absence of evidence as evidence of absence (i.e., that they
will internalize a lack of positive evidence as amounting to negative evidence), particularly consider- ing that a grammatical version of the sentence exists in English with two possible readings. In theory, if English speakers’ dominant language influences their interpretations, and if no one ever informs them that the interpretation with the indirect object as an antecedent is impossible in Japanese, the fact that a sentence has never been used with that interpretation may not weaken the alternative reading to the extent that they reject it completely. In such a situation, negative evidence would seem to be necessary.

It is important to keep in mind that the results of empirical studies do not consistently indicate that the L1 influences learners’ interpretations of reflexives in an L2 (see Chapter 3). Nevertheless, theoretically at least, this type of situation could be seen as provisionally contributing to making restrictions on antecedents for *zibun* viable as a linguistic target for testing S. E. Carroll’s (2001, 2007) ‘awareness constraint on negative evidence’. As discussed in Chapter 1, based on arguments regarding the architecture of the language faculty and the aspects of language assumed to be accessible to awareness, Carroll claims that it should be impossible for learners to benefit from corrective feedback or metalinguistic information on linguistic phenomena for which relevant distinctions are not somehow represented phonologically. Syntactically, reflexives arguably fulfill this criterion since conditions on their interpretation involve the structural relationships of c-command and subjecthood, neither of which is associated unambiguously with a particular surface form. Overt morphosyntactic clues, such as nominative case-marking or occurrence in sentence-initial position, are not trustworthy indicators of subject status or of c-command in Japanese since both subjects and objects can be marked nominative or dative (depending on the verb, for example), and since c-commanding
and non-c-commanding elements can appear anywhere in the sentence, linearly speaking. As explained in more detail below, this was exploited in the testing and treatment materials to ensure that, in order to interpret zibun in a reliably targetlike manner, learners had to be sensitive to features of underlying hierarchical relationships which, according to Carroll, should be inaccessible to awareness. Moreover, as also discussed in Chapter 3, since there are many known exceptions to the generalization that reflexives’ binding behavior is related to their morphological form, proposals adverting to morphological triggers of UG seem problematic.

Considering some of the apparent contradictions in the findings of CALL experiments comparing more and less explicit feedback types (see Chapter 1), another major criterion in selecting reflexives as a linguistic target in this study was that metalinguistic information should be able to impart notions that learners might not come up with on their own. In some studies (e.g., Camblor-Portilla, 2006; Moreno, 2007; Sanz & Morgan-Short, 2004), learners receiving feedback informing them only of the correctness/incorrectness of their answers have shown levels of improvement similar to those in more explicit feedback groups. In such cases, the researchers have not tended to argue that generalizations emerged bottom-up for learners in the less explicit conditions, but rather that learners in those conditions were likely able to figure out the rules without the help of additional information. As Sanz and Morgan-Short (2004) put it, “when there are a limited number of possibilities in the input and the number of responses is also limited, locating the source of the problem and rejecting hypotheses is possible even without explicit feedback” (p. 73). Thus, it seemed reasonable to select an area of language where many of the sentences would be complex, binary feedback would not be exhaustive (i.e., knowing that something was right would not necessarily entail knowing that something else
was wrong, and vice versa), systematicity might not be readily apparent, and blame assignment might be difficult. Although at least one relevant rule for \textit{zibun}'s interpretation can be stated rather straightforwardly (i.e., that an antecedent must be a c-commanding subject), the learning task is complicated by the wide variety of sometimes ambiguous sentences in which \textit{zibun} can appear, by the fact that most learners are not familiar with the concept of c-command, and by the fact that grammatical subjects might not always be obviously recognizable as such (e.g., in ECM constructions and causatives, such as \textit{Koizumi-ga Bussyu-ni zibun}_{j} o syokaisuru yooni settokusita} ['Koizumi persuaded Bush_{j} to introduce self_{j}'] and \textit{Baato-ga Miruhausuj-ni zibun}_{j} no syukudai-o saseta} ['Barti made Milhouse_{j} do self_{j}’s homework'], in which Bush and Milhouse are dative-marked subjects of the embedded clauses and can serve as antecedents of \textit{zibun}.

In relation to questions of aptitude, which this study also sought to investigate, several of the above factors (e.g., the multiplicity of interpretations and importance of abstract metalinguistic information, the necessity of tracking referents in complex sentences) would seem to increase the likelihood that learners with different profiles of abilities and knowledge (e.g., in sensitivity to ambiguity, grammatical sensitivity, familiarity with metalinguistic analyses) might be more or less equipped to benefit from different types of information.

Finally, researchers have verified that this is an area of language where learners do not receive explicit instruction in the classroom. Thomas (1991), for example, examined several Japanese language textbooks and found no evidence that learners are taught about constraints on interpretations of \textit{zibun}. When Thomas (1995) interviewed participants at the end of another study, none claimed to have learned previously about \textit{zibun}'s relevant properties;
typically, they responded that they had learned it as a vocabulary item and did not “recall anything other than the general meaning of ‘self’ or ‘one’s own’” (p. 224). Informal interviews with Japanese instructors at the various institutions where data were collected for the current study (e.g., Y. Mori and M. Hama, Georgetown University; M. Mason, University of Maryland) confirmed that *zibun* tends to be presented as a lexical item meaning ‘oneself’ and that more specific details on its uses are treated only in courses on Japanese linguistics.

**Overview of design**

The study involved a computer-mediated experiment conducted over five sessions in a pre-test/post-test/delayed-post-test design with random assignment to groups. A diagram summarizing the procedure is shown in Figure 4.2. Following a brief general overview here, each step and measurement instrument will be discussed in more detail below.

On the first day, I went through the informed consent process with each participant and, after consent had been obtained, asked him or her to complete the following activities: an initial (pre-actional) motivation questionnaire, a test of sensitivity to ambiguity and familiarity with metalinguistic terminology in English, a pre-test of the ability to interpret Japanese sentences containing the reflexive *zibun*, another (actional) motivation questionnaire following the pre-test, and a background/biodata questionnaire.

The second session began with the first experimental treatment. Each participant was given a PowerPoint file designed to provide practice in interpreting sentences with *zibun*. In accordance with the participant’s random group assignment, the hyperlinks in the PowerPoint were equipped either (1) to indicate whether or not each sentence interpretation was correct
(Right/Wrong Feedback), (2) to indicate whether or not each interpretation was correct and to illustrate the relevant linguistic structure with phrase-structure tree diagrams (Trees Feedback), or (3) to provide no feedback (No Feedback). Also by random assignment, the participant either was or was not asked to produce a concurrent verbal protocol. Instructions were given both verbally by the researcher and in written form within each PowerPoint file. Upon completion of the first treatment activity, each participant again filled out an actional motivation questionnaire. Considering how much time was available in the participant’s schedule that day, s/he was then given a choice of individual-difference measures to complete: the Modern Language Aptitude Test (MLAT), the Visual Patterns Test (VPT), an operation-span test (Ospan), and/or a listening-span test (Lspan) of working memory capacity. The third session was conducted in the same manner.

The fourth experimental session likewise began with a treatment activity followed by a motivation questionnaire, but this time, instead of completing an individual-difference measure afterwards, the participant was given an immediate post-test of his/her ability to interpret Japanese sentences with zibun. This was followed by a reflection questionnaire asking what the participant thought s/he had noticed and learned about Japanese, as well as a final (post-actional) motivation questionnaire. Except in a few cases when the participants’ availability did not allow it, the three treatments (Sessions 2-4) were scheduled within a two-week period. On average, the number of days from Treatment 1 to Treatment 3 and the immediate post-test was 10.3. Broken down by experimental group, it was 10.5 days for the Trees condition, 10.2 days for the Right/Wrong condition, and 10.4 days for the No-Feedback condition.
At the fifth session, which took place approximately a week after the immediate post-test, each participant took a delayed post-test of his/her ability to interpret Japanese sentences with *zibun* and completed any remaining individual-difference measures (unless there were more than two measures remaining, in which case an additional session was planned to complete them). At the end of each participant’s last session, I provided an individualized verbal explanation of the linguistic target, walked the participant through several examples to illustrate how the rules worked, described the differences between the experimental groups, and answered any questions the participant posed regarding the experiment or second language acquisition in general.

The success of the random assignment is discussed in Chapter 5 (Results), where Table 5.6 displays the mean scores of the three experimental groups on each individual-difference (ID) measure (e.g., length of Japanese study, grammatical sensitivity, visual short-term memory, etc.). In short, as explained in that chapter, a series of omnibus one-way ANOVAs, run separately for each ID, found no statistically significant differences among the treatment groups. This can increase our confidence that they were sufficiently equivalent in terms of background characteristics and cognitive abilities for valid comparisons among them to be made.
Figure 4.2. Overview of the research procedure

**Session 1** (Week 1)
1. Informed consent
2. Motivation questionnaire (pre-actional)
3. Test of sensitivity to ambiguity and knowledge of metalinguistic terminology
4. Pre-test of interpreting Japanese sentences with zibun (counterbalanced)
   - Version A/A2 (n=29)
   - Version B/B2 (n=25)
   - Version C/C2 (n=26)
5. Motivation questionnaire (actional)
6. Background questionnaire

**Sessions 2, 3, and 4** (Weeks 2-3)
(Participants completed each of the following activities at each session.)
1. Treatment
   - No Feedback (n=25)
     - TA (n=12)
     - Non-TA (n=13)
   - Right/Wrong Feedback (n=29)
     - TA (n=15)
     - Non-TA (n=14)
   - Trees Feedback (n=26)
     - TA (n=13)
     - Non-TA (n=13)
2. Motivation questionnaire (actional)

  **Sessions 2 and 3:**
3. Individual-difference measure (given choice of MLAT, VPT, Ospan, &/or Lspan)

  **Session 4:**
3. Immediate post-test of interpreting Japanese sentences with zibun
   - Version A/A2 (n=24)
   - Version B/B2 (n=27)
   - Version C/C2 (n=29)
4. Reflection questionnaire
5. Motivation questionnaire (post-actional)

**Session 5** (Week 4)
1. Delayed post-test of interpreting Japanese sentences with zibun
   - Version A/A2 (n=27)
   - Version B/B2 (n=28)
   - Version C/C2 (n=25)
2. Individual-difference measures (any remaining)
3. Explanation from the researcher
Recruitment of volunteers and the informed consent process

All aspects of this research project were officially approved by Georgetown University’s Institutional Review Board (IRB) before recruiting or meeting with any participants. The first step in the recruitment process was to email the director of the university’s Japanese language program, a professor of Japanese listed as an appropriate contact on the program’s web site, or a professor who had personally offered to facilitate the recruiting process. The next step, once permission had been granted, was to email individual professors who were teaching Japanese at the intermediate level or above at each university. After explaining my research goals, I asked whether they would mind letting me visit their classes in order to let their students know about the opportunity to participate in the study.

During each recruiting session, which lasted 10-15 minutes, I handed out information sheets describing the goals of the study (see Appendix A). I began by explaining that I was a graduate student at Georgetown University working on a Ph.D. in applied linguistics, specifically in the area of adult second language acquisition. I mentioned that I was visiting their class to ask whether anyone would be interested in participating in my dissertation research, which I described as having three main goals. First, I explained that many linguists were interested in trying to predict specifically what sorts of challenges people with a particular language background (like English) might encounter in trying to learn a particular second language (like Japanese). I mentioned that I had done some research into the similarities and differences between English and Japanese and had decided to focus on an area of Japanese grammar where some interesting predictions could be made. Thus, the first goal of my research, as I explained it, was to investigate what sorts of information could help Japanese learners to
improve in that area of grammar. Second, I noted that, being interested in language pedagogy, another important question for me was whether one method of helping learners with this grammar point would work better than another; thus, I had designed a series of computer-based activities in order to test different approaches to practicing it. I described the activities as involving sentences illustrated with funny pictures of celebrities and cartoon characters. I also mentioned that participants would have a chance to practice their *katakana* skills (since foreign names are transliterated using that syllabary) and to expand their knowledge of *kanji* (since glosses would sometimes be provided in English and/or *hiragana*, a more often practiced syllabary). Third, I brought up the issue of individual differences. I mentioned that research had shown that instructional techniques that work quite well for one person might not be as effective for another, and I gave as an example that some people might tend to be more memory- or communication-oriented, whereas others might prefer analyzing grammar and trying to figure out sets of rules. Thus, in my experiment, I would also be measuring a variety of cognitive differences, such working memory capacity and grammatical sensitivity, to see whether individual characteristics might interact with how effective each experimental condition was. In other words, without informing the students about the specific area of Japanese grammar being targeted, I was fairly transparent about the goals and design of the research, and this may have had an influence on their motivation, expectations, and reasons for participating.

Once I had finished explaining the three themes of my study, I told the prospective volunteers that participating would involve meeting with me five times over the course of a few weeks in a computer lab on their campus and that I would be glad to meet with them at any
time that fit with their schedules. I handed out and collected sign-up sheets, and after leaving, I contacted the volunteers individually by email or phone, according to their expressed preferences.

At my first meeting with each participant, I began by walking him or her through the informed consent document (see Appendix B) and offering to answer any questions s/he might have. I made it very clear to the participants that they would be randomly assigned to one of three possible experimental conditions. I admitted that one (or more) might work better than the other(s) and pointed out that there was no guarantee that they would be matched with learning activities that fit best with their personal characteristics. However, I also mentioned that I would explain everything about the linguistic target, cognitive measures, and research design at the end of the experiment, as well as answer any remaining questions they might have. I asked that they please not discuss the experiment with any of the other participants until everyone had completely finished all of the activities, and I explained that all of their data would be kept confidential: Each participant would be assigned a number, and all of the documents and audio files produced over the course of the experiment would be associated with that number. Finally, I emphasized that participation was entirely voluntary at all times. All participants were given a copy of the informed consent document detailing this information.

The next sections describe the tests and treatments on interpretations of *zibun*, including their basic format; how they were designed, piloted, administered, and scored; and how the concurrent think-aloud protocols were conducted. Then, the focus turns to the individual-difference measures: the background questionnaire, the test of sensitivity to ambiguity and familiarity with metalinguistic terminology, the Modern Language Aptitude Test
(J. B. Carroll & Sapon, 1959), and the Visual Patterns Test (Della Sala, Gray, Baddeley, & Wilson, 1997). The data from the motivation questionnaires, reflection questionnaire, and working memory tests (listening span and operation span) have not yet been analyzed; thus, they are not discussed here. However, descriptions and answer sheets are available in Appendices G1-G5 (motivation and reflection) and H1-H2 (working memory).

**Design of the tests and treatments**

**Truth-value judgment task**

A notorious methodological problem in empirical studies of both native speakers’ and L2 learners’ interpretations of reflexives is that preferences for certain interpretations over others can play a role in how participants respond. That is, even when a sentence is ambiguous and a person’s grammar allows more than one possible interpretation, a primary interpretation may tend to win out; speakers’ habitual preferences can “suppress [their] perception of underlying ambiguity” (Thomas, 1991, pp. 224-5) and block or override their recognition of secondary interpretations. As MacLaughlin (1998) puts it,

[L]inguistically-naïve participants... independent of whether they are tested on their native language or on a second language—are notoriously poor at reporting multiple interpretations. Even when they receive training to detect ambiguity, or when the experimental design is set up to present learners directly with multiple interpretations... learners still seem unable to recognize alternative interpretations (p. 205).
A major goal in the design of the tests and treatments for the current experiment was to avoid this problem, to the extent possible. In the literature on L2 reflexives, some task types have been found to be more successful at this than others. In general, studies employing picture-identification tasks (e.g., Christie, 1992; Finer, 1991; Finer & Broselow, 1986) and multiple-choice interpretation tasks (Hirakawa, 1990; Matsumura, 1994; Thomas, 1989, 1991; Yuan, 1998) might be considered “contaminated” by the preference effect (Matsumura, 2007). However, studies employing truth-value judgment tasks (e.g., Akiyama, 2002; Hamilton, 1998; L. Jiang, 2009; Matsumura, 2007; Thomas, 1995; Wells, 1998; White, 1995b; White et al., 1996) may be less so.

In the mid-1990s, dissatisfied with the questionable validity of picture-identification and multiple-choice interpretation task formats, several researchers (e.g., Thomas, 1995; White, 1995b; White et al., 1996) began to turn away from them and toward written versions of a truth-value judgment task (TVJT) which had been developed by Stephen Crain and colleagues (e.g., Crain & McKee, 1985; Crain & Fodor, 1989; Crain & Thornton, 1998) to investigate child L1 acquisition. In that area as well, researchers had encountered the problem of interpretation preferences and had found themselves misled into underestimating children’s knowledge. To address this problem, they developed a task in which an experimenter would use toys or puppets to act out a scenario, following which another toy or puppet would make a comment, and the experimenter would ask the child whether the comment was true or not based on what had just happened.

In TVJTs, according to Bruhn-Garavito (1995), “a certain interpretation is forced on the subject, who is then free to reject it if it is not valid” (p. 85). White et al. (1997) point out that
since the targeted interpretation is highlighted within a clear context, participants are never required to consider more than one interpretation at once, and “the problem of a potential preference for an alternative interpretation [does] not arise” (p. 149). Some of the sentences used in these tasks are ambiguous and some are not, but, importantly, the question is simply whether or not each can be interpreted in a given way. Thus, from White’s (1995b) perspective, “this task is a major improvement over traditional judgment tasks, since it allows one to probe the learner’s unconscious knowledge... without asking for an explicit judgment of the form of a sentence” (p. 76). In principle, participants are asked only about truth or falsity, not for any metalinguistic analysis. Another advantage, noted by Slabakova (2008), is that the expected answer is “categorically True or False, and never both” (p. 124). A review of research methods employed in studies of L2 reflexives is provided in Appendix C.

In the present experiment, both the tests and the treatments took the form of truth-value judgment tasks, using both written narrative contexts and pictorial representations (Thomas, 1995; White et al., 1997) as a way of overcoming some of the methodological drawbacks encountered in previous studies. All of the sentences employed were grammatical, and the participants were simply asked whether or not each target sentence could be used in the given narrative context to express the meaning that was highlighted by the picture. The word zibun appeared only in the target sentences whose truth values the participants were to judge, never in the narrative contexts. Despite the fact that zibun is not marked for gender, all of the people and/or cartoon characters mentioned in each sentence were of the same gender (or, if animals were involved, it could be assumed that they were of the same gender). This was done for two reasons: (1) because third-person singular reflexives in English (e.g., himself,
herself) indicate the gender of the antecedent, and, hypothetically, that might influence interpretations in the L2, and (2) because long-distance binding (for Chinese ziji, at least) may be possible only when the person and number features of the potential antecedents in a sentence match, even if the reflexive itself does not mark such features.\textsuperscript{29} Great care was also taken to try to ensure that all of the characters in each sentence would seem equally plausible as antecedents. In other words, the participants should not have been overly influenced by their real-world knowledge or expectations in deciding whether an antecedent was allowable; their choices should have been due solely to the grammatical structure of each sentence.

The TVJT\textsc{s} were created in PowerPoint, using pictures found through \textit{Google Images} (http://www.google.com/imghp) and \textit{Microsoft Office Online} (http://office.microsoft.com/en-us/images). Each slide began with a contextualizing sentence in Japanese whose intended function was to bias the participants toward a particular interpretation of the target sentence that would soon be appearing. After 3 seconds, a picture appeared, with the target sentence below it. Then, after 3 more seconds, the target sentence was underlined, and the question “Can the sentence be used in this situation?” appeared (in Japanese) along with two response buttons, labeled \textit{Hai} (‘yes’) and \textit{Iie} (‘no’). Underneath this question, in parentheses, a brief phrase clarified further the interpretation that was being targeted. The response buttons, which simulated button-presses when clicked, were augmented with hyperlinks. During the

\textsuperscript{29} For example, compare sentences (a) and (b), reported in Beck (1998, p. 8):

(a) \textit{Zhangsan, renwei [Lisi, zhidaow Wangwu, xihuan ziji\_\texti/\textj]}  
‘Zhangsan thinks that Lisi knows that Wangwu likes himself’

(b) \textit{Zhangsan, renwei [wo, haile ziji\_\texti/\textj]}  
‘Zhangsan knows that I hurt myself’

In (a), any of the people mentioned can serve as an antecedent for \textit{ziji}. In (b), only the speaker can.
tests, they led participants to the next slide. During the treatments, according to experimental
group assignment, they led participants to different slides depending on how they answered.

This basic format of the TVJT is illustrated in Figure 4.3, which shows two different items
associated with the sentence *Inu-ga neko-ni zibun-no tabemono-o tottekosaseto* (‘The dog
made the cat fetch self’s food’). In the top slide pictured, which tests whether the cat can be an
antecedent for *zibun*, the context sentence says that dogs do not enjoy touching cats’ food, and
the parenthetical underneath the question “Can the sentence be used...’ reads “i.e., the cat’s
food.” In the bottom slide, which tests whether the dog can be an antecedent for *zibun*, the
context sentence says that the dog was hungry, but he was injured, and the parenthetical reads
“i.e., the dog’s food.” Since this is a causative structure, and both the dog and the cat are
considered to be subjects which c-command *zibun* (Tsujimura, 2007), both of them are
theoretically possible antecedents. As such, clicking ‘yes’ would be considered correct in both
cases. For both the tests and the treatments, in addition to clicking the buttons in the
PowerPoint, the participants were asked to record their answers in writing on paper answer
sheets (see Appendices D1-D2).

During the tests, as shown in Figure 4.4, immediately after responding yes or no for each
item, there was also a slide asking for the participant’s assessment of his or her level of
confidence in making the judgment as well as for an indication of what the response had been
based on: a guess, intuition, memory, or a rule. This was inspired by arguments in the implicit
learning literature to the effect that such confidence ratings and source attributions can shed
important additional light on the types of knowledge learners have developed, beyond what
researchers can gather from verbal reports on debriefing questionnaires (see, e.g., Rebuschat &
Williams, 2009). The data obtained from the confidence ratings and source attributions in the current study are not reported in this dissertation. However, future analyses are planned.

**Figure 4.3.** Example of two items associated with a causative sentence
The instructions for the tests and treatments are provided in full in Appendices E1 and E2, respectively. Before taking any of the tests, the participants were informed of the format that the items would take. It was emphasized that all of the sentences would be grammatical in Japanese and that the participants should simply decide whether or not the sentence at the bottom of each slide could be used with the meaning implied by the context. An animated example was provided in order to walk them through the steps for responding, using a target sentence with the same interpretation in English and Japanese. Then, the next two slides explained how to go about making confidence judgments and source attributions. The participants were allowed to ask the researcher questions about unfamiliar vocabulary, but they were asked not to return to any previous items or to change any of their answers.
The instructions for the treatments differed according to experimental condition. For instance, whereas participants in the No-Feedback condition were simply encouraged to continue paying attention to the meanings of the sentences and to try to have fun, those in the Right/Wrong- and Trees-Feedback conditions were additionally informed that the computer would be telling them whether their judgments were right or wrong. In all three groups, however, it was noted that the activity would be similar to the previous one (i.e., the pre-test, or the previous treatment), highlighting the same points about grammaticality and a focus on meaning. During the treatments, the participants were not asked to provide confidence ratings or source attributions.

Treatment conditions

Figures 4.5, 4.6, and 4.7 illustrate what happened during the treatment activities for the No-Feedback, Right/Wrong-Feedback, and Trees-Feedback groups, respectively. In the No-Feedback treatment condition, the participants engaged in practice interpreting sentences containing zibun, but were never informed about the accuracy of their responses. Instead, at various points over the course of each activity, as shown in Figure 4.5, brief expressions of encouragement were interspersed among the items (e.g., asking the participants if they were having fun, joking about the people featured in the sentences, letting them know that they were almost done, etc.). In the Right/Wrong-Feedback condition, the participants likewise engaged in practice interpreting sentences containing zibun. In contrast to the No-Feedback condition, however, immediately after clicking ‘yes’ or ‘no’ for each item, they were presented with a slide which informed them whether their answer had been right or wrong. As shown in
Figure 4.6, each of these feedback slides displayed a happy- or sad-looking cartoon character (either Pikachu, a popular creature from the Pokémon franchise created by Satoshi Tajiri; Doraemon, a robotic cat from a Japanese manga series created by Fujiko Fujio; Totoro, a character in a Japanese anime film by Hayao Miyazaki; or Hello Kitty, a cartoon cat designed by Yuko Shimizu and produced by the Japanese company Sanrio). Alongside the picture, there was a feedback message which began with a short positive or negative phrase (e.g., *Yatta!* [Yay!], *Atari!* [Right on target/you got it!], *Seekai!* [Correct!], *Pinpoon!* [onomatopoeia indicating that something is correct]; or *Zannen...* [Sorry], *Hazure...* [You didn’t get it], *Huseekai...* [Incorrect], *Matigae...* [Mistake]), followed by an explicit statement regarding whether or not the sentence could be used in the given situation.

In the Trees-Feedback condition, in addition to engaging in practice interpreting sentences containing *zibun* and being informed immediately afterward whether each response had been right or wrong, the participants were presented with an animated tree diagram whenever they responded incorrectly. As shown in Figure 4.7, this tree diagram illustrated not only the overall structure of the sentence, but also the structural relationship between *zibun* and whichever potential antecedent was being targeted for that item. When the tree diagram first appeared, the slide showed only a skeleton of the sentence structure with the nodes’ grammatical categories labeled. Then, part by part, the words of the sentence appeared. For instance, for the sentence shown in Figure 4.7, *inu-ga* (‘dog’ + nominative) appeared first at the top of the diagram, followed by *neko-ni* (‘cat’ + dative), then the words *zibun-no tabemono-o* (‘self’ + genitive, ‘food’ + accusative), followed by the verb *totteko* (‘fetch’) and the causative *saseta* (‘make’ in the past tense). The amount of time it took for each to appear ranged from 1
to 3 seconds, depending on how many words were involved. Once all of the words were in place, the antecedent being targeted turned bright blue and flashed; then, an arrow wiped upwards across the screen to the S node, which likewise turned greenish-blue and flashed as a circle appeared around it. Next, arrows started wiping downwards, one by one, until they reached zibun. At this point, zibun turned bright blue and flashed as a circle appeared around it, and a smiley-face pinwheeled in to indicate that the antecedent was, indeed, possible. When the participants were ready to move on to the next sentence, they clicked the arrow at the bottom-right corner of the screen.

As part of the treatment instructions, the participants in the Trees condition were given some very basic information about how tree diagrams work. This is shown in Figure 4.8. The first slide introduced the idea of tree diagrams along with an example of a simple sentence: *Chuck-ga inu-o ketta* (‘Chuck kicked the dog’). The next slide was animated to facilitate an explanation of each part of the diagram in turn. As a description appeared on the right side of the screen, the relevant grammatical category was highlighted and circled in the diagram. The following slide presented a diagram for a slightly more complicated sentence: *Chuck-ga ketta inu-ga zibun-o kiratta* (‘The dog that Chuck kicked hated himself’). Next, this diagram was deconstructed, also in animated fashion. Again using blue circles, the participants were pointed to what they might understand as the main subject of the sentence (‘the dog’ [that Chuck kicked]); to the relative clause modifying it (‘Chuck kicked’); to the verb phrase serving as the predicate of the sentence (‘hated himself’); and to the fact that the phrase *Chuck-ga ketta inu-ga* (‘the dog that Chuck kicked’) as a whole made up its own noun phrase. Finally, the participants were given the opportunity to replay the instructions or to move on. Importantly,
at no point during the instructions were they given any indications as to which antecedent(s) *zibun* could or could not refer to.

*Figure 4.5. Examples of encouragement slides in the No-Feedback condition*

Translations of the Japanese text on each slide: (A) Are you having fun? Totoro is having fun! Next is Dick Cheney. Are you ready? (B) Hello Kitty is impressed by your diligence. (C) Now Pikachu is singing for you! (D) Keep it up! The end is in sight!
Figure 4.6. Examples of the feedback provided in the Right/Wrong condition

Translators of the Japanese text on each slide: (A) Yay! The sentence can be used in this situation. (Literally: Under the conditions of the picture, this sentence is possible to use.) (B) You didn’t get it... Actually, the sentence can be used in this situation. (C) [Onomatopoeia indicating the answer was correct] The sentence cannot be used in this situation. (D) It’s not allowed! Actually, the sentence cannot be used in this situation.
Figure 4.7. Diagrams provided in the Trees condition for the causative sentence shown in Figure 4.3.
Linguists have a variety of ways of describing and analyzing the structure of sentences. One way is by using tree diagrams.

So, for example, you could represent the structure of a sentence like チャックが犬をけった ('Chuck kicked the dog') as follows:

```
S
  |   |   |
N V  |
  |   |   |
N V  |
  |   |   |
チャックが 犬を けった
```

Let’s look at the various parts of this diagram in a bit more detail…

A sentence or clause is labeled S…

Nouns are labeled N…

Verbs are labeled V…

and nodes that join together more than one element are labeled as phrases, like the VP (Verb Phrase) here.

チャックが is the subject of the sentence. What did he do? 犬をけった。

The words in the VP describe the whole action (what happened and who or what was affected by it).
Of course, ideas and sentences can get more complicated…

For example:

チャックがけった犬が自分をきらった。

Again, we can find the main subject of the sentence: in this case, the noun 犬。

This time, 犬 is described by a whole clause: the S encompassing チャックが犬をけった。

And what did the 犬 do (or, in this case, feel)? Check out the VP (Verb Phrase):自分をきらった。
You may also have noticed that がけった犬が itself makes up a larger Noun Phrase (NP).

Which dog (N) are we talking about in the rest of the sentence? The dog that Chuck kicked (NP).

Well, that’s about it! This may take a little while to get used to, but you’ll have plenty of practice! 😊

**Sentence types used in the tests and treatments**

As explained above, one of the goals of the current study was to test S. E. Carroll’s (2001) claim that corrective feedback and metalinguistic information cannot lead to learning in cases where relevant distinctions are not manifested in surface representations. Since subjects in Japanese canonically occur in sentence-initial position and are often marked with nominative case (-ga), the treatment and testing materials had to be manipulated in such a way that any rules of thumb the participants might develop based on overt morphological and/or syntactic clues would not consistently produce targetlike interpretations. Fortunately, the canonical situation in Japanese is not the only possible situation. It is also possible for c-commanding subjects to be marked with particles other than –ga (e.g., the embedded subjects in causatives...
and the dative subjects of certain verbs, which are marked –ni), for non-c-commanding subjects to be marked with –ga (e.g., in relative clauses), and for non-subjects to be marked with –ga (e.g., as the objects of certain verbs). In the latter two cases, regardless of case-marking, non-c-commanding subjects and non-subjects cannot serve as antecedents for zibun. Moreover, since relative clauses are prenominal in Japanese, sentences can begin with non-c-commanding subjects, marked either nominative or dative.

Table 4.2 illustrates how various characteristics were balanced against each other within each test and treatment. Each test was designed to include sentences checking the participants’ acceptance of local and long-distance c-commanding subjects as antecedents for zibun, as well as their rejection of non-subjects and non-c-commanding subjects. Translation equivalents of the items testing for acceptance of local c-commanding subjects and rejection of non-c-commanding subjects were considered to work the same way in English and Japanese in the sense that the same antecedents would be allowed and disallowed in the two languages. Items testing for acceptance of long-distance c-commanding subjects as antecedents for zibun, on the other hand, might be hypothesized to require positive evidence for English-speaking learners of L2 Japanese, whereas items testing for rejection of non-subject antecedents might be hypothesized to require negative evidence. On each test, the number of items theoretically requiring positive evidence was equal to the number theoretically requiring negative evidence (with 9 of each), while 12 items theoretically should have worked the same in English and Japanese. Moreover, each of these item types was balanced as equally as possible across the halves of the test, with 4 or 5 positive-evidence items on each half, 4 or 5 negative-evidence items on each half, and 6 items that worked the same way in English on each half. The numbers
of correct ‘yes’ and ‘no’ responses were also nearly balanced, with 16 ‘yes’ responses (8 on each half) and 14 ‘no’ responses (7 on each half).

To undercut potential learner-generated hypotheses based on surface characteristics, the number of targeted antecedents marked with nominative –\textit{ga} (16) was roughly equal to the number marked with dative –\textit{ni} or accusative –\textit{o} (10+4 = 14). For those antecedents marked –\textit{ga}, the ratio of correct ‘yes’ to correct ‘no’ responses was 9 to 7, while for –\textit{ni} and –\textit{o} it was 7 to 7. Concerning the sentence positions of the targeted antecedents, the first noun was targeted 15 times, and the second or third noun in the sentence was targeted the other 15 times. A complete list of the test sentences, broken down by sentence type, is provided in Appendix F1, along with brief descriptions of the types of antecedents targeted on each test version.

Like the tests, the treatments were designed to include sentences focusing on the acceptance of local and long-distance c-commanding subjects as antecedents for \textit{zibun}, as well as the rejection of non-subjects and non-c-commanding subjects as antecedents. Thus, in theory, both positive evidence and negative evidence were provided when feedback was given. As shown in Table 4.2, each treatment contained between 24-26 items which were considered to function the same way in English and Japanese as far as allowable antecedents were concerned (75 total), and these were split roughly equally across the halves of each treatment. Between 6-8 items theoretically involved positive evidence (21 total), and between 7-9 items theoretically involved negative evidence (24 total). Regarding the number of correct ‘yes’ and ‘no’ responses, these were also balanced evenly across the halves of each treatment (between 11-13 ‘yes’ responses per half, and 7-9 ‘no’ responses per half); however, there was a greater
number of items whose answers were ‘yes’ in each treatment (23-24) compared to the number of items whose answers were ‘no’ (16-17).

As far as case morphology is concerned, no items in any of the treatments contained antecedents marked with accusative case (-o); 23-24 items per treatment involved nominative-marked antecedents (-ga), while 16-17 involved dative-marked antecedents (-ni). Within each of these categories, there were more ‘yes’ responses than ‘no’ responses: 13-15 ‘yes’ versus 9-11 ‘no’ responses for antecedents marked nominative -ga, and 9-10 ‘yes’ versus 6-7 ‘no’ responses for antecedents marked dative -ni. With regard to sentence complexity, Treatment 1 was intended to involve relatively simple structures to start out; thus, none of the items involved sentences with more than 2 targeted antecedents. The first noun in the sentence was targeted 17 times, while the second noun was targeted 23 times. In Treatments 2 and 3, in which sentences with up to 3 targeted antecedents were included, the first noun was targeted 15-18 times, the second noun 18-19 times, and a third noun 3-7 times. A complete list of the treatment sentences, broken down by sentence type, is provided in Appendix F2, along with brief descriptions of the antecedents targeted in each session.

Each treatment activity consisted of 40 items, presented in exactly the same order for all participants, regardless of experimental condition. Thus, across the 3 treatment sessions, each participant judged 120 target sentences, and all of the participants judged all the same sentences. In most cases, each sentence was included in a treatment session as many times as there were nouns whose status as an antecedent could be tested. In other words, if a sentence had 2 potential antecedents, it usually appeared 2 times during the same treatment, once in the first half and once in the second half, testing a different antecedent each time. If a sentence
had 3 potential antecedents, it appeared 3 times, once in each third of the treatment. For the few sentences which were included in a treatment only once, this was done with the purpose of balancing the characteristics outlined in Table 4.2.

In deciding how many items to use, concerns about participant fatigue (Gass & Mackey, 2007) were weighed against the possibility that including fewer items might bias results in favor of the most explicit treatment (DeKeyser, 2003; N. Ellis, 1993, 2005; R. Ellis, 2009; Long, 2007). As Long (2007) has pointed out, there is an “often unrecognized built-in advantage” for explicit instruction in SLA research. “Incidental and implicit learning,” he notes, “typically require longer to work” (p. 123). Due to logistical constraints, however, SLA researchers often perform short-term laboratory studies. In the CALL feedback studies reviewed in Chapter 1, the number of treatment items ranged from 16 (Moreno, 2007) to 112 (Lado, 2008). The number of items in the present study therefore exceeded the numbers used in previous research, although, of course, this is no guarantee of avoiding an explicit-learning bias; it was nowhere near the 15 hours of self-study performed by participants in de Graaff’s (1997) experiment comparing more versus less explicit instruction, which was discussed in Chapter 2 in relation to aptitude.
Table 4.2. Overview of test- and treatment-item characteristics

<table>
<thead>
<tr>
<th></th>
<th>TESTS (30 items)</th>
<th>TREATMENTS (40 items each, 120 total)</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentence types</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same as English</td>
<td>12 (6/6)</td>
<td>25 (12/13)</td>
<td>24 (13/11)</td>
<td>26 (14/12)</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>Positive evidence</td>
<td>9 (5/4)</td>
<td>6 (4/2)</td>
<td>8 (4/4)</td>
<td>7 (4/3)</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Negative evidence</td>
<td>9 (4/5)</td>
<td>9 (4/5)</td>
<td>8 (3/5)</td>
<td>7 (2/5)</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td><strong>Yes &amp; no responses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>16 (8/8)</td>
<td>24 (12/12)</td>
<td>24 (11/13)</td>
<td>23 (12/11)</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>14 (7/7)</td>
<td>16 (8/8)</td>
<td>16 (9/7)</td>
<td>17 (8/9)</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td><strong>Case morphology of targeted nouns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GA (nominative)</td>
<td>16 (9 yes 7 no)</td>
<td>24 (15 yes 9 no)</td>
<td>23 (14 yes 9 no)</td>
<td>24 (13 yes 11 no)</td>
<td>71 (42 yes 29 no)</td>
<td></td>
</tr>
<tr>
<td>NI (dative)</td>
<td>10 (6 yes 4 no)</td>
<td>16 (9 yes 7 no)</td>
<td>17 (10 yes 7 no)</td>
<td>16 (10 yes 6 no)</td>
<td>49 (35 yes 20 no)</td>
<td></td>
</tr>
<tr>
<td>O (accusative)</td>
<td>4 (1 yes 3 no)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>Position in sentence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noun 1</td>
<td>15 (10 ga 5 ni)</td>
<td>17 (12 ga 5 ni)</td>
<td>15 (9 ga 6 ni)</td>
<td>18 (12 ga 6 ni)</td>
<td>50 (33 ga 17 ni)</td>
<td></td>
</tr>
<tr>
<td>Noun 2</td>
<td>9 (4 ga 3 ni 2 o)</td>
<td>23 (12 ga 11 ni)</td>
<td>18 (12 ga 6 ni)</td>
<td>19 (11 ga 8 ni)</td>
<td>60 (35 ga 25 ni)</td>
<td></td>
</tr>
<tr>
<td>Noun 3</td>
<td>6 (1 ga 2 ni 2 o)</td>
<td>0</td>
<td>7 (2 ga 5 ni)</td>
<td>3 (1 ga 2 ni)</td>
<td>10 (3 ga 7 ni)</td>
<td></td>
</tr>
</tbody>
</table>
Counterbalancing of test forms

The tests of learners’ interpretations of *zibun* were created in 3 (or, in a sense, 6) parallel forms (Versions A/A2, B/B2, and C/C2), which were counterbalanced across the pre-test, post-test, and delayed post-test in case there turned out to be differences in the difficulty of the versions. The difference between Versions A/A2 vs. B/B2 vs. C/C2 was in the specific sentences that were used; exactly the same sentence *types* appeared on each test version so that the same types of antecedents could be tested in the same ways across versions. The difference between Versions A vs. A2, Version B vs. B2, and Versions C vs. C2 was in the ordering of the items. In order to try to control for fatigue (as advised by Gass & Mackey, 2007), items which had occurred toward the beginning of Version X occurred toward the end of Version X2, and vice versa, for each pair. (More specifically, divided into quadrants, items 1-7, 8-14, 15-21, and 22-28 on Version X became items 15-21, 22-28, 1-7, and 8-14 on Version X2, respectively, and the ordering of the items within each quadrant was reversed so that, for example, items 1, 2, 3, 4, 5, 6, and 7 became items 21, 20, 19, 18, 17, 16, and 15.)

Another difference between the X and X2 versions, admittedly less than ideal, was that 2 additional items were added to each X2 version (as item #8 and as item #23 or #26). This was done in order to be able to include another type of generalization item which the participants had not practiced during the treatments (specifically, biclausal nonfinite structures), the theoretical importance of which became clear after data collection had already begun. Thus, each X version had 28 items, whereas the corresponding X2 version had exactly those same 28 in a different order, plus 2 more, for a total of 30 items. For 31 participants, Versions A, B, and C were counterbalanced; for the other 49 participants, Versions A2, B2, and C2 were.
Looking only at pre-test scores, evidence was not found that the test versions differed from one another; a one-way ANOVA with Version as the between-groups variable and pre-test scores as the dependent variable did not produce a statistically significant result ($F(5, 74) = 1.872, p = 0.11$). Looking only at post-test scores, the outcome was similar ($F(5, 73) = 2.139, p = 0.07$), although at $p = 0.07$ the result approached statistical significance. However, looking only at delayed post-test scores, a one-way ANOVA comparing the test versions did show a statistically significant difference ($F(5, 78) = 4.729, p = .001^{**}$). Scheffé post-hoc tests indicated that performance on Version B (77%) was significantly different from performance on Versions A2 (62%), B2 (60%), and C2 (61%).

How can this be explained, considering that the items on Versions B and B2 were exactly the same apart from the 2 extra items added to Version B2? As will be discussed in greater detail in Chapter 5 (Results), it may have had to do with the abilities of the participants who took the X versions versus those who took the X2 versions. All of the participants who took the X versions during the Spring, Summer, and Fall semesters of 2009 came from two universities with intensive Japanese language courses, and the average delayed post-test scores on each of those versions ranged from 67-77%. The participants who took the X2 versions during the Fall semester of 2009 and the Spring semester of 2010 included students from two other universities as well, and the average delayed post-test scores on each of those versions ranged from 60-62%. In other words, although it might look as though the X2 versions were more difficult, it may actually be that the participants who took the X versions had higher ability.
Piloting of the testing and treatment items

The development and piloting of the testing and treatment items was an iterative, multi-step process and a collaborative effort throughout. Several fellow graduate students in Georgetown University’s Linguistics Department, all native speakers of Japanese who had taught or were currently teaching Japanese classes in Georgetown’s Department of East Asian Languages and Cultures, provided extensive assistance in materials development. One worked with me to create viable Japanese versions of target sentences that I had attempted to create based on my reading of the literature in Japanese linguistics. Another worked with me to transfer all of the sentences into Japanese orthography and to create the context sentences that would be used to introduce each target sentence. Two others checked all of the PowerPoint slides to ensure not only that they were accurate, but also that learners of Japanese at an intermediate level or above could be expected to know the vocabulary words and kanji (logographs) employed. In cases where this seemed questionable, they suggested alternate words or provided glosses in English or hiragana.

All of the test items were then piloted by several adult native speakers of Japanese (3 men and 3 women)—friends of mine who had not been involved in the item-development process and did not know the purposes of the study. All were from Japan originally, but most (5/6) had completed undergraduate or graduate degrees in the United States or Canada as international students; the last was also highly proficient in English, having taken classes at the Advanced Academic level in an intensive English program.

First, 3 of these Japanese speakers piloted rough drafts of all 3 test versions, scoring 22-25 out of 28 (79-89%), with an average score of 23 (83%). After reviewing the specific items
where their responses had not matched what was intended, and considering the comments they had made regarding sources of confusion, I worked with one of my Georgetown colleagues to revise the contextualizing sentences, pictures, and/or target sentences, as relevant. Then, the next 3 Japanese speakers piloted all 3 tests, scoring 25-28 out of 28 (89-100%), with an average score of 26 (93%). Across all of the versions, only 1 item was missed by more than 1 person during this second round of piloting, and it involved a local subject in a causative structure, which could be expected to have a weak reading for native speakers of Japanese in any case (Akiyama, 2002; Fukui, 1984; Hirakawa, 1990; White et al., 1997).30

Once all 6 of the native speakers of Japanese had piloted all of the tests for the first time, the items were revised further. Several of the context sentences were modified in an attempt to ensure that inferences regarding who was supposed to be doing what to whom were straightforward and pragmatically appropriate. For instance, for an item targeting a (disallowed) dative indirect object as an antecedent for *zibun* (*Kanzya-ga isya-ni zibun-nituite kiita*, ‘The patient asked the doctor about himself’), the context had originally read that patients sometimes do not trust doctors, with the supposed implication being that patients might therefore want to ask their doctors questions. However, this inference was not necessarily

---

30 According to Fukui (1984), long-distance antecedents for *zibun* are “obviously dominant” while local antecedents are “very weak” (p. 23, cited by Thomas, 1991, p. 225, note 30), and it has frequently been found that native speakers of Japanese may “fail to recognize” that local subjects can serve as antecedents for *zibun* in sentences where long-distance antecedents are also available (White et al., 1997, p. 148). In Hirakawa’s (1990) study, 63% of the native speakers of Japanese bound *zibun* exclusively to long-distance antecedents in sentences like *Taro-wa Akira-ga zibun-o butta-to itta* (‘Taro said that Akira hit self’) (p. 71). In White et al.’s (1996) study, 5 of the 9 native speakers of Japanese consistently rejected local subject antecedents in biclausal sentences, whereas the other 4 consistently accepted them. In light of this finding, and considering theoretical proposals by Hyams and Sigurjónsdóttir (1990) regarding Korean long-distance reflexive *caki*, they speculate as one possibility that Japanese speakers might fall into two different dialect groups (p. 250). Noting a similar problem, Akiyama (2002) points out that “it is not always easy to obtain the ambiguous interpretation that potentially exists in biclausal Japanese sentences” since it may depend on a variety of factors, including choices of verbs and case markers (p. 36). Moreover, for causatives in particular, pragmatic conditions can play a large role in determining whether a given sentence is appropriate to suggest that someone is making another person do something.
obvious; thus, the context sentence was revised to say more explicitly that patients often want to know more information about their doctors.

Another problematic issue was that the pictures were not always as clear as they could have been at first. Sometimes, the pilot testers mentioned that they had not been able to recognize all of the American celebrities (e.g., Hugh Hefner), cartoon characters (e.g., Liono from Thundercats), or politicians (e.g., Joseph Biden), or that it had simply not been obvious which character in a picture was being referred to. To address these problems, I labeled many more of the pictures with the characters’ names, included arrows to indicate the directionality of various actions (e.g., giving, sending, kicking), added icons to suggest people’s emotions more clearly (e.g., depictions of smiling, crying, or angry faces), and tried to make the targeted antecedent more salient in the picture (e.g., by increasing the size of the image).

A problem related to the linguistic form of the target sentences which arose repeatedly was that some of the pilot testers felt very uncomfortable with sentences containing more than one nominative marking (-ga). Sentences containing more than one –ga are, strictly speaking, grammatical; however, they are uncommon and somewhat awkward in Japanese. In Akiyama’s (2002) study as well, some of the participants “found the sequence of NP–ga NP–ga awkward and so presumably placed primacy on the first NP” (p. 50, note 15). For the sake of naturalness, it would have been preferable to use –wa (a topic marker) for one of the nouns marked nominative in each double-ga sentence; however, considering the purposes of the experiment, it seemed crucial to avoid this. For one thing, topicalization might have served to highlight one antecedent over another; for another, it would have complicated the process of balancing the sentences’ morphological characteristics. Thus, in order to make these target items feel more
natural, we changed the context sentences to the form of questions asking who had done something. In Japanese, such questions require nominative subjects (not topics) in response.

After these revisions, I compiled subsets of items that at least 1 pilot tester had answered in an unexpected way, then sent each pilot tester an individualized subset of items. This subset included all of the sentences s/he had answered in an unexpected way, plus some of the sentences s/he had answered in the expected way. In this round of piloting, 2 of the pilot testers continued to answer 2-4 items in an unexpected way on each version, rejecting antecedents s/he might have been expected to accept (e.g., local subjects with weak readings); however, 3 of the pilot testers answered all of the items correctly.

For the treatment items, the piloting process was similar in most respects to that described above for the tests, the main difference being that the scheduling was a bit different: First, a rough draft was piloted with 5 native speakers, who scored around 80% on average; then, a revised version was piloted with a single native speaker. Again, in most cases, if any of the pilot testers answered an item in an unexpected way, only 1 or 2 of them did so. The items which were answered in an unexpected way by more than 2 pilot testers most often involved local subjects as antecedents. Following an analysis of the pilot testers’ response patterns and comments, the treatment items were revised yet again in the same manner that the test items had been revised.

As White et al. (1996) point out, “performance at 100% accuracy by native speakers is unusual in any experimental attempts to get at linguistic competence; investigation of binding theory is no exception” (p. 249, note 10). However, particularly considering the theoretical controversies over the status of zibun (see Appendix F3) and the fact that some less common
sentence types were employed in this experiment in an attempt to balance various
morphological and syntactic characteristics, it seemed very important to take seriously the goal
of creating items that native speakers of Japanese would judge as expected.

Test scoring and reliability

In scoring the final versions of the tests which the L2 learners took during the
experiment itself, 1 point was awarded for each correct answer and 0 points for each incorrect
answer. To assess the internal consistency (reliability) of each test version, Cronbach’s alpha (α)
was employed, including just the 28 items which had appeared on both the X and X2 versions of
each test form. A commonly cited rule of thumb in the social sciences is that a test can be
considered sufficiently reliable if the value of α is at least .70 (Larson-Hall, 2010, p. 171). Overall,
none of the versions of the tests (A/A2, B/B2, or C/C2) reached this level, ranging instead from
α=.59 to α=.64. In interpreting the results of this study, therefore, it will be important to keep in
mind that the tests were less reliable than might have been desired, and that noise in the data
may have worked against finding treatment effects or finding clear relationships between
individual differences and L2 performance.

Reporting on similarly low reliability coefficients for some of the tests employed in her
study, Erlam (2005) speculates that they may have been due to the tests’ difficulty levels or
“may suggest that students were guessing answers” (p. 156). De Graaff (1997) likewise explains
a Kuder-Richardson-20 reliability coefficient of .59 on one of his tests (equivalent to Cronbach’s
α for dichotomous items) as being “due to performance on chance level for a considerable
number of participants” (p. 260). As we will see in the next chapter, learners in the
Right/Wrong- and No-Feedback groups, considered independently, did not show statistically significant improvement over the course of the tests; they started out on the pre-test scoring around 59% on average, and by the time of the post-tests were scoring only around 62-65% on average. If we consider, in light of these test scores, that they may have been behaving more randomly than learners in the Trees-Feedback condition, who scored 74% on the post-test on average, then it might make sense to find differences in the reliability coefficients among the groups, and, indeed, that is what we find. As illustrated in Table 4.3, when Cronbach’s α is calculated for the 3 experimental groups separately, the tests do appear to have been more reliable for the Trees-Feedback group (.66 ≤ α ≤ .71), whereas they were less reliable for the Right/Wrong-Feedback group (.44 ≤ α ≤ .59) and the No-Feedback group (.41 ≤ α ≤ .66).31

Table 4.3. Cronbach’s α for each test version across experimental groups

<table>
<thead>
<tr>
<th>Version A/A2</th>
<th>Version B/B2</th>
<th>Version C/C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>All groups</td>
<td>.61</td>
<td>.59</td>
</tr>
<tr>
<td>Trees:</td>
<td>.70</td>
<td>.66</td>
</tr>
<tr>
<td>Right/Wrong:</td>
<td>.59</td>
<td>.52</td>
</tr>
<tr>
<td>No Feedback:</td>
<td>.41^a</td>
<td>.49</td>
</tr>
</tbody>
</table>

^a Two items involving local nominative-marked subjects as antecedents were deleted by the statistical program due to zero variance: *Eimii-ga Kerii-o zibun-no ie-no mae-de matta* (‘Amy waited for Kelly outside self’s house’) and *Obaasan-ga onnanoko-ni zibun-nituite-no hanasi-o sita* (‘The grandmother told the girl a story about herself’).

^b One item, also involving a local nominative-marked subject, was deleted by the statistical program due to zero variance: *Taanya-ga obaasan-o zibun-no ie-ni tureteitta* (‘Tanya led the old woman to self’s house’).

31 Another perspective on reliability can be gained by calculating Cronbach’s α separately for subscales of items intended to test different theoretical constructs (e.g., negative versus positive evidence here). When this was done, however, the reliabilities of the subscales were not usually found to be higher than those for the instruments as a whole, whether considering all of the participants together or considering the experimental groups separately.
In order to determine whether any particular test items might have been functioning inconsistently and might therefore have been candidates for exclusion, item analyses were carried out. On each test version, several items were identified whose exclusion would have increased the overall consistency of the instrument; however, they tended to show up in pairs targeting the same antecedent type. For instance, on Version A/A2, the items whose exclusion would have increased the value of $\alpha$ included 2 items testing the acceptance of long-distance dative-marked subject antecedents (positive evidence), 2 testing the acceptance of local dative- or accusative-marked causative subject antecedents (similar to English), 2 testing the rejection of local nominative-marked object antecedents (negative evidence), and 1 testing the rejection of a local accusative-marked object antecedent (negative evidence). Excluding all of these items would have meant excluding 25% of the items on the instrument, and in many cases would have meant excluding not just one malfunctioning item among those intended to test a particular antecedent type, but rather excluding most or all of the items for that antecedent type. In relation to the goal of ensuring that participants would not be able to achieve perfect scores through employing hypotheses based on surface characteristics, this would have been undesirable. Nearly all of the items which appeared to function inconsistently involved non-canonical case marking (i.e., dative- or accusative-marked subjects and nominative-marked objects).

Moreover, the items which functioned inconsistently for the participants in one experimental condition did not always seem to function inconsistently for the participants in another. For example, considering the Trees group’s performance independently, the items whose exclusion would have increased the internal consistency of Version A/A2 included 1 of
the previously mentioned items testing LD dative subject antecedents, both of the local causative subjects, and 2 of the local objects (1 nominative, 1 accusative), but not the other LD dative subject or the other local nominative object. On top of these, 2 additional items emerged as candidates for exclusion. Considering the Right/Wrong group independently, the items whose exclusion would have increased the internal consistency of Version A/A2 included 1 of the local causative subjects and both local nominative objects, but neither of the LD dative subjects, nor the other causative subject or local accusative object. On top of these, 4 additional items emerged as candidates for exclusion. Considering the No-Feedback group independently, the items whose exclusion would have increased the internal consistency of Version A/A2 included both of the LD dative subjects, both of the causative subjects, and the local accusative object, but neither of the local nominative objects. On top of these, 7 additional items emerged as candidates for exclusion. In view of these discrepancies, the decision was made not to exclude any items from the analyses presented here. However, it may be a productive direction for more detailed future analyses to take.

Finally, a word about the exclusion of participants based on their test scores: In many experimental studies of SLA, participants are excluded for performing at a level above chance on the pre-test. Since each item on the tests of the present study involved a binary choice, theoretically, a participant could have achieved a score of 50% purely by chance. However, considering that 12/30 [i.e., 40%] of the test items were hypothesized to work the same way in English and Japanese, it seemed plausible that participants might be able to answer those correctly at the outset, and considering that the participants may already have been exposed to positive evidence of long-distance binding in previous L2 input, it seemed plausible that they
might be able to answer some of those items correctly as well. Therefore, the threshold for excluding participants was set at a score of 80% on the pre-test, reasoning that this cut-off would still leave room for improvement and avoid ceiling effects.

**Concurrent verbal protocols**

Many researchers in the field of SLA (e.g., Alanen, 1995; Bowles & Leow, 2005; Camps, 2003; Leow, 1997, 1998, 2000, 2001; Leow, Egi, Nuevo, & Tsai, 2003; Rosa & Leow, 2004a; Rosa & O’Neill, 1999; Sachs & Polio, 2007; Sachs & Suh, 2007; Sanz, Lin, Lado, Bowden, & Stafford, 2009; Swain & Lapkin, 1995) have found participants’ concurrent verbalizations to be a rich source of insights regarding L2 learners’ allocation of cognitive resources, uses of strategies, and awareness of specific L2 forms and feedback. Introspective data can reveal information about learners’ approaches to L2 tasks that would not be available otherwise (Gass & Mackey, 2000). It has even been argued (e.g., by Leow, 1999) that fine-grained online measures are necessary if SLA researchers wish to move beyond merely speculating about learners’ attentional processes. Concurrent think-aloud protocols were therefore included in the design of the present study under the rationale that introspective data would ultimately be necessary to address the assumptions underlying Carroll’s (2001) awareness constraint on negative evidence. If the results are taken to suggest that learners can benefit from visual feedback on reflexive binding, then measures of awareness will be required to take the extra step of ascertaining why Carroll’s prediction appears to be wrong. If the results do not suggest this, then measures of awareness might help to explain why her prediction appears to be right.
Concerns have repeatedly been raised, however, that uses of think-aloud protocols as a research tool might be reactive—that is, that they might change the very thought processes that SLA researchers are trying to measure (see, e.g., Bowles, 2010; Cohen, 2000; R. Ellis, 2001; Jourdenais, 2001; Wigglesworth, 2005). As such, an increasing number of studies have been examining empirically whether concurrent verbalization exerts either detrimental or beneficial effects on L2 comprehension, production, and learning (e.g., Bowles, 2005, 2008; Bowles & Leow, 2005; Goo, 2010; Leow & Morgan-Short, 2004; Polio & Wang, 2005; Rossomondo, 2007; Sachs & Polio, 2007; Sachs & Suh, 2007; Sanz et al., 2009). The results have been mixed, and it has accordingly been cautioned that researchers should include a silent group as a control whenever think-alouds are used as a research tool (Leow & Morgan-Short, 2004; Sanz et al., 2009). That advice was followed here: Approximately half of the participants in each treatment condition were asked to think aloud during the treatments, whereas the other half were asked to complete them silently.32 Statistical analyses were then conducted, as a methodological check, to determine whether verbalizing concurrently might have affected the amount of improvement the participants demonstrated over the course of the tests.

In Ericsson and Simon’s (1993) comprehensive review of the use of verbal reports in non-SLA fields, and in Bowles’ (2010) book on the think-aloud controversy in L2 research, the authors explain that different methods of eliciting verbalizations may affect the data obtained. Ericsson and Simon emphasize, for instance, the importance of not asking participants for justifications or explanations of what they are thinking. The instructions for thinking aloud (available in Appendix E3) therefore began by referring to the idea of inner speech and

32 In the No-Feedback group, 12 learners thought aloud while 13 did not; in the Right/Wrong-Feedback group, 15 thought aloud while 14 did not; and in the Trees-Feedback group, 13 thought aloud while 13 did not.
explained to the participants that they should not worry about trying to describe their thoughts to anyone else or trying to provide reasons for why they were thinking what they were thinking; rather, the idea was simply to vocalize whatever happened to be going through their minds at the time. The participants were assured that everyone else in the room, including the researcher, would be wearing either earplugs or headphones except when it was necessary for the researcher to give instructions or answer questions. It was also emphasized that thinking aloud was not a test and that they should not worry about whether what they were saying was correct or incorrect, or even about whether they were speaking in complete sentences. The participants were encouraged to speak in whichever language happened to be going through their minds and felt most comfortable at the moment. In order to help them get acclimated to speaking their thoughts aloud before beginning the Japanese task, they were asked to verbalize concurrently while solving a multiplication problem on a sheet of scrap paper (as done by Bowles & Leow, 2005; Leow & Morgan-Short, 2004; Sanz et al., 2009; see also Ericsson & Simon, 1993). Finally, they were given the opportunity to ask any questions that they might have about the think-aloud process and were reminded to keep speaking throughout the whole activity, simply vocalizing whatever entered their minds. The verbalizations were audio-recorded in different ways according to the equipment available at each research site. In some cases, this meant using recording software installed on the computers themselves (e.g., Audacity) and headphones with built-in microphones. In other cases, it meant using handheld digital voice recorders (e.g., an Olympus WS-321M).

Although more detailed analyses can certainly be pursued (e.g., in relation to individual differences), a basic answer to the question of reactivity in this experiment is straightforward:
Over the course of the tests, the Think-Aloud (TA) and Non-Think-Aloud (Non-TA) groups never showed statistically significant differences from each other with regard to their accuracy in interpreting sentences with reflexives.\(^{33}\) A repeated-measures analysis of variance (RM ANOVA) comparing the groups’ performance over time, with Think-Aloud group as the between-subjects factor and Time (pre, post, delay) as the within-subjects factor, showed a statistically significant overall effect for Time \((F(1.75, 133.31) = 14.667^*, p < 0.001, \eta_p^2 = 0.16)\), indicating that the participants improved over the course of the tests in general; however, it produced neither a statistically significant Think-Aloud effect \((F(1, 76) = 0.093, p = 0.76, \eta_p^2 = 0.001)\) nor a statistically significant Time-by-Think-Aloud interaction effect \((F(1.75, 133.31) = 2.081, p = 0.14, \eta_p^2 = 0.03)\). (The former would have indicated that the groups differed from each other overall, and the latter would have indicated a difference in their trajectories of test performance over time.) According to another set of RM ANOVAs, conducted separately for each Think-Aloud condition, both the TA participants \((F(1.71, 64.99) = 12.907^*, p < 0.001, \eta_p^2 = 0.25)\) and the Non-TA participants \((F(1.71, 64.84) = 4.162^*, p = 0.03, \eta_p^2 = 0.10)\) demonstrated statistically significant improvement. A statistically non-significant difference of roughly 3% in their mean pre-test scores \((t(76) = 1.083, p = 0.28, d = 0.25)\) became statistically non-significant differences of 1% on the post-test \((t(76) = -0.224, p = 0.82, d = 0.05)\) and 4% on the delayed post-test \((t(76) = -1.395, p = 0.17, d = 0.03)\). Thus, it does not appear to be the case that thinking aloud was reactive.

---

\(^{33}\) Although the fact seems often to be unrecognized or underemphasized, if differences in participants’ accuracy of test performance are not found, that cannot be taken as a definitive signal that their thought processes during the treatments were not altered by the verbalization requirement. After all, it is possible to achieve similar test scores via more than one approach to learning. However, an apparent lack of a reactivity effect on the accuracy of participants’ test performance at least does not give researchers an unmistakable reason to worry that verbalizing affected the participants’ thought processes.
Appendix L. Information regarding the amount of time participants spent during the treatments in the Think-Aloud group versus in the Non-Think-Aloud group can be found in Appendix M.

*Individual-difference measures*

**Background questionnaire**

The background questionnaire, which was completed by the participants at the end of the first experimental session, was designed to elicit information regarding their previous study of foreign languages, current uses of Japanese, and experience with linguistics, as well as basic biodata such as age, gender, and native language(s). The participants were asked which Japanese course(s) they were currently enrolled in, how long they had been studying Japanese and in what contexts, why they were studying Japanese, whether they had spent any time living or traveling in Japan, how many hours per week they used Japanese outside the classroom and for what purposes, what other languages they were familiar with at what levels of proficiency, and how they had learned those other languages. They were also asked about their academic majors, which linguistics courses they had taken (if any), whether they had ever learned to diagram sentences, how easy diagramming had been (as relevant), and whether or not they enjoyed grammar. On the last two questions, they were asked to elaborate, and their more detailed responses were coded on 5-point Likert-style scales ranging from difficult to easy and from emphatic hate to emphatic love. A full version of the background questionnaire is provided in Appendix K.
Test of sensitivity to ambiguity and knowledge of metalinguistic terminology

Ambiguity section

It is not only in an L2 that variability can be found in people’s sensitivity to linguistic ambiguity; in fact, there is variation among individuals in their L1 as well. MacLaughlin (1998), for example, reports that some of the native-speaker controls in her study were “much better at detecting and reporting ambiguity than others” (p. 221, note 13). Because of this, several researchers in the area of L2 reflexives have included pre-task training sessions to try to ensure that their participants would approach the experimental tasks with an awareness of the fact that ambiguities can exist (e.g., Thomas, 1989, 1991, 1995; White et al., 1997; Yip & Tang, 1998). In Thomas’ (1989, 1991) studies of L2 reflexives, she trained her participants first on sentences with pronouns, such as John thinks that he can find out whether Bill passed the test (where ‘he’ can refer only to John) and John thinks that Bill can find out whether he passed the test (where ‘he’ can refer to John or Bill). To prepare her participants for a story-based task type which she used in a later study of L2 Japanese, Thomas (1995) had them practice the task beforehand by reading a story in English and deciding whether two follow-up statements accurately reflected the content of the story: ‘A’ thinks that he is getting sick and He thinks that ‘A’ is getting sick. Yip and Tang (1998) attempted to train their participants using ambiguous sentences which were unrelated to reflexives, such as John saw Bill drunk.

In the current study, immediately after filling out the pre-actional motivation questionnaire and before taking the pre-test of zibun interpretation, the participants completed an activity consisting of 2 subtests: a measure of their sensitivity to ambiguity in English (discussed here) and a measure of their knowledge of metalinguistic terminology (to be
discussed below). The purpose of the ambiguity measure was twofold: (1) as a baseline individual-difference measure of sensitivity to ambiguity that was at least relatively unaffected by L2 proficiency, and (2) as a means of heightening the participants’ awareness of the fact that it was possible to have clear intuitions about whether or not a given sentence had more than one possible reading. The latter was related to the methodological concern that their responses on the experimental tasks should reflect their L2 grammars as accurately as possible. The answer sheets for the test of sensitivity to ambiguity and knowledge of metalinguistic terminology are provided in Appendices H1-H2.

**Design**

The ambiguity subtest contained 16 items, 8 of which involved reflexives and 8 of which involved other linguistic phenomena. The items with reflexives tested 4 different issues, with 2 items devoted to each: long-distance antecedents with infinitival/small clauses (e.g., *Yoda told Luke to trust himself*), long-distance antecedents in causatives (e.g., *The pitching coach made the baseball player watch himself*), non-subject antecedents (e.g., *The amnesic asked the psychologist some questions about himself*), and non-c-commanding antecedents (e.g., *The professor who saw the student cheating blamed himself*). One of the items which tested for a non-c-commanding antecedent additionally tested for a long-distance antecedent across a finite clause boundary (*The homecoming queen that the mean cheerleader punched suspected that the valedictorian hated herself*). Importantly, the genders of the people mentioned in these items were kept constant (or could be interpreted as constant) within each sentence. The items testing sensitivity to ambiguity with linguistic phenomena other than reflexives involved 5
different issues: strict vs. sloppy readings under ellipsis (e.g., *Joe loves his mother, and so does Jim*), relative-clause and/or prepositional-phrase attachment (e.g., *The podiatrist envied the shoe salesman with two assistants who loved feet; Picasso discussed painting with Dalí*), and scope of negation and/or quantifiers (e.g., *It didn’t snow on more than 2 days; A girl kissed every boy*). The different types of sentences were distributed over the test in such a way that sentences of the same type did not immediately follow one another.

The instructions, which were presented to the participants in written form and also explained verbally, began with an example asking them to consider the sentence *We need more intelligent teachers.* Its 2 possible interpretations were indicated, and the participants were informed that some of the sentences they were about to see would be ambiguous, like the example, whereas others would be unambiguous. Their task was to “paraphrase the meaning(s) of each sentence in a very clear, unambiguous way.” Later, during the instructions for the Japanese pre-test, it was mentioned that, although the format would be different, the Japanese activity would be similar, in a sense, to the ambiguity activity they had completed in English: On the Japanese test, instead of being asked to generate the possible meaning(s) of sentences themselves, they would simply be asked whether a particular interpretation of each sentence was possible. Sometimes, they might recognize that a sentence could have more than one meaning, but each question would always ask about only one interpretation, and they should judge each item only with respect to the meaning that was being targeted.
Scoring and reliability

In scoring the ambiguity subtest, participants were awarded 1 point for each item they answered in the intended way. Since one of the items tested for the rejection of two different types of antecedent, the 16 items were worth 17 points in total. Of these 17 points, 11 were considered to involve recognizing an ambiguity, while 6 involved effectively rejecting a disallowed antecedent (i.e., not reporting a non-existent ambiguity, whether by reporting only one possible interpretation or by reporting only those interpretations that were actually possible in English). For instance, in the sentence *The pitching coach made the baseball player watch himself*, a participant would receive 1 point for describing only a situation in which the baseball player (the local subject) was watching himself. If a participant described a situation in which the baseball player was watching the pitching coach (a disallowed long-distance antecedent), 0 points would be awarded. However, paraphrases involving lexical ambiguities were not scored, so, for example, if a participant reported that the sentence could involve both a literal sort of watching (e.g., on video) and a figurative sort of watching (e.g., being careful), s/he could still score 1 point for that sentence. If an item was supposed to be ambiguous, on the other hand, 1 point was awarded for reporting more than 1 possible reading, again related to the intended type of ambiguity. For instance, in the sentence *Nancy Pelosi read Sarah Palin a book about herself*, participants would receive 1 point for reporting that the book could be about either Nancy or Sarah. However, if a participant paraphrased the sentence as meaning only that the book was about Nancy, 0 points were awarded. For example, one participant reported that the sentence could mean that Nancy read a book about herself to Sarah and that “the book Nancy read was called ‘Sarah Palin’ and it was about herself (Nancy).” Since the
participant did not recognize the non-subject antecedent (Sarah), 0 points were awarded. In a few cases, when participants seemed to be trying to describe 2 possible readings, but did not describe them completely accurately, a half point was occasionally awarded. For instance, one participant was given a half point for paraphrasing the sentence *A girl kissed every boy* as meaning that there was one girl who kissed every boy and that each single girl kissed every boy. (Similar examples included “each girl gave all the boys one kiss; one girl gave all the boys a kiss” and “each girl kissed a boy; one girl kissed all the boys.”)

In order to ensure that this section of the test was scored in a reliable manner, inter-rater agreement was checked with another Georgetown Linguistics graduate student, who rated a portion of the data independently. As a first step, the second rater took the test herself, as though she were a research participant, and then compared her answers against an answer key which I provided. In our first meeting, we reviewed the test items, discussed the coding scheme, and worked through scoring one participant’s test together. At the end of the meeting, I emailed her 4 more participants’ tests to score independently (i.e., roughly 5% of the data). Once she had rated them on her own, we met again to check our codes against one another’s, identify the reasons for any discrepancies, and refine the coding scheme as necessary, modifying it or clarifying how particular types of responses should be scored. Then, I sent her 9 more participants’ tests to score independently (i.e., roughly 11% of the data), randomly selecting 1 or 2 from each ‘decade’ of participant numbers (i.e., 1 whose participant number was in the 20s, 1 from the 30s, 1 from the 40s, etc.). Our rate of agreement on the data from these 9 participants was 97%, with a Cohen’s Kappa value of $\kappa = .90$ (qualifying as ‘almost perfect’ agreement, according to Landis and Koch, 1977). The 5 discrepancies, out of 153
possible items, were due to 2 simple mistakes, 1 actual difference in judgment, and 2
differences related to unexpected participant behavior which had not been built into the coding
scheme. After discussing these discrepancies and refining the coding scheme again where
necessary, I scored the rest of the ambiguity data myself. The participants’ scores ranged from
5.5 (32.4%) to 17 (100%), with a mean of 12.5 (73.3%) and a standard deviation of 3.0 (17.4%).
The internal consistency of the measure, as assessed by Cronbach’s alpha, was $\alpha=.73$. As such,
it was considered sufficiently reliable (Larson-Hall, 2010, p. 171).

**Metalinguistic terminology section**

As discussed in Chapter 2, several studies (e.g., Alderson et al., 1997; Berry, 2009) have
found that university students vary widely in their knowledge of terms for grammatical
constructs. In the context of the present experiment, this type of knowledge was considered
potentially relevant in light of the fact that some participants would be provided with abstract
and highly metalinguistic feedback in a possibly unfamiliar format regarding a complex linguistic
target for which they would not have been taught any rules. Additionally, in light of the fact
that the participants in the Think-Aloud condition would be required to verbalize their thought
processes, it seemed important to obtain at least some indication of their familiarity with
formal grammatical constructs and the possession of terminology which might facilitate their
expression. Therefore, as part of the same document as the test of sensitivity to ambiguity, the
participants were given a brief test of their knowledge of metalinguistic terminology.
**Design**

The contents and formats of several other researchers’ tests were consulted in designing the measure of familiarity with metalinguistic terminology for the current study. One early metalinguistic knowledge test, designed by Alderson et al. (1997), was composed of 125 items and took 50 minutes to complete. In Section 1, which had 4 parts, the participants were asked to read sentences in either their L1, English (in Parts 1 and 2), or their L2, French (in Parts 3 and 4), and identify various parts of speech. In Sections 2 (French) and 3 (English), they were asked to read sentences, each of which contained a grammatical error which was highlighted, then correct the error and state a relevant rule. Many other researchers have since modeled their tests on Alderson et al.’s (1997) instrument.

In Roehr and Gánem-Gutiérrez’s (2009) experiment, for example, a somewhat similar test of metalinguistic knowledge was given just in the learners’ L2. It included 2 sections. Section 1 contained 20 items, each of which presented an L2 sentence with an error highlighted. The participants’ task was to correct, describe, and explain the error. Section 2 contained 25 items modeled on Part IV of the Modern Language Aptitude Test and was designed to test the participants’ ability to recognize the grammatical roles of words in L2 sentences. As in Section 1, part of each sentence was highlighted; however, in this section, “[n]o description, explanation, explicit labeling, or use of technical terminology was required” (p. 169). The participants simply had to choose from among 4 options the word or phrase “which they regarded as playing an analogous grammatical role” (p. 170).

A metalinguistic knowledge test whose validation is discussed by Elder (2009) also included 2 sections. Section 1, with 17 items, involved choosing from several multiple-choice
options which rule explanation was the most appropriate to describe an error that had been underlined in an ungrammatical L2 sentence. Section 2, with 19 items, involved “match[ing] items from a list of grammatical terms to their corresponding exemplars in an [L2] sentence” (p. 118). Elder notes that both parts of the test involved some degree of familiarity with metalinguistic terms, but she also points out that the type of metalinguistic knowledge tested was passive in the sense that it did not involve the ability to generate L2 rules independently.

Berry (2009) gave his participants a 50-item test including metalinguistic terms in the L2 “ranging from basic word classes (noun, verb, etc.) and verb forms (present tense, infinitive, etc.) to more esoteric terms from scientific grammar (predicate, complement, etc.)” (p. 115). The participants were asked to demonstrate their knowledge of each by providing a stand-alone example or by underlining the relevant word(s) in a phrase or clause which they themselves had written.

The test of familiarity with metalinguistic terminology employed in the present study was perhaps most similar to Alderson et al.’s (1997) Section 1, Parts 1 and 2, which involved identifying parts of speech in the L1. The participants were presented with a single paragraph, 164 words in length, which they were asked to read and then use as a source of examples for 16 grammatical terms. Several of the terms were potentially relevant to stating rules about reflexives (e.g., direct object, indirect object, reflexive pronoun, antecedent, relative clause, causative), whereas others were not obviously related (e.g., subordinating conjunction, adverbial subordinate clause, interrogative nominal clause, appositive). None of them were simple parts of speech (noun, verb) or verb tenses (present, past); thus, overall, the content might be considered more difficult than that of the tests used in the studies by Alderson et al.
(1997) and Berry (2009), which did include those more commonly known terms. Taking into account the large battery of other activities the participants in this experiment would also be completing, it was decided to limit the scope of the metalinguistic knowledge test so that it could be administered in a relatively short time and would be relatively independent of L2 proficiency. It did not assess the ability to correct errors or to recognize or state specific rules about the L2 (as in Alderson et al., 1997; Elder, 2009; Roehr & Gánem-Gutiérrez, 2009). It also did not directly test sensitivity to the roles of words in sentences (as in Roehr & Gánem-Gutiérrez, 2009), as that would be evaluated separately on the Modern Language Aptitude Test.

**Scoring and reliability**

In scoring the test of familiarity with metalinguistic terminology, participants were awarded 1 point for identifying an appropriate example from the paragraph and 0 points for leaving an item blank or responding with a word or phrase that was not actually an example of the metalinguistic term. A half point was awarded for a response which was partially accurate in the sense that its characteristics matched at least one aspect of the grammatical construct being targeted. For example, if a participant used an adverb as an example of an adverbial subordinate clause, s/he would be given partial credit for responding with something adverbal; if a participant used a restrictive relative clause as an example of a non-restrictive relative clause, s/he would be given partial credit for identifying a relative clause.

Exactly the same procedure for assessing inter-rater agreement was followed as that described above for the test of sensitivity to ambiguity. After refining the coding system, our rate of agreement on 9 participants’ tests was 94% (135/144), with a Kappa value of $\kappa = .90$. 

238
The discrepancies involved 3 simple mistakes, 1 misunderstanding of the technical definition of one of the grammatical terms, and 5 disagreements in which we had reasoned similarly but awarded partial credit differently. After reaching agreements on all of the items we had coded differently, the coding scheme was revised further, and I scored the rest of the data. The participants’ scores on the test of familiarity with metalinguistic terminology ranged from 0 to 14.5 (90.6%), with a mean of 6.3 (39.2%) and a standard deviation of 3.5 (21.6%). The internal consistency of the instrument, as assessed by Cronbach’s alpha, was $\alpha=.86$.

**The Modern Language Aptitude Test**

As explained in Chapter 2, The Modern Language Aptitude Test, or MLAT (J.B. Carroll & Sapon, 1959), has formed the “cornerstone of aptitude research” since the mid-twentieth century (Skehan, 2002). It has been validated for a variety of literate, English-speaking adult populations, including university students, and a significant amount of aptitude research has continued to follow the agenda set by its developers. According to the test manual (J.B. Carroll & Sapon, 2002, p. 3), the MLAT’s five subtests can be said to measure the following constructs:

- **Part I (Number Learning):** “one aspect of the memory component of foreign language aptitude... [and possibly] a special ‘auditory alertness’ factor which would play a role in auditory comprehension of a foreign language”

- **Part II (Phonetic Script):** “the ability to learn correspondences between speech sounds and orthographic symbols... [and possibly] memory for speech sounds”
• Part III (Spelling Clues): “English vocabulary knowledge… [and] the same kind of sound-symbol association ability as measured by Part II… highly speeded”

• Part IV (Words in Sentences): “sensitivity to grammatical structure, and [possibly]… the student’s ability to handle the grammatical aspects of a foreign language…. [N]o grammatical terminology is involved, so that the scores do not depend upon specific memory for grammatical terminology”

• Part V (Paired Associates): “the rote memory aspect of the learning of foreign languages”

Among the other major aptitude tests which have been developed, the PLAB (Pimsleur’s Language Aptitude Battery, 1966), DLAB (Defense Language Aptitude Battery, Petersen & Al-Haik, 1976), and VORD (Parry & Child, 1990) show greater emphasis on inductive language leaning ability, which seems important for the purposes of the current experiment. However, the DLAB and VORD have not been found to demonstrate higher predictive validity than the MLAT (Skehan, 2002), and the MLAT is commercially available through Second Language Testing, Inc. (a company based in Rockville, Maryland), whereas the DLAB is restricted to U.S. government use. As for the PLAB, considering that it was designed to be used in grades 7 through 12, and considering that its language analysis section tests the ability to extract rules related to case-markings and word order, it was thought that it might be too easy or familiar for the university learners of L2 Japanese who participated in the current study.
The Short Form of the MLAT includes only Parts III (sound-symbol associations), IV (grammatical sensitivity), and V (rote memory for linguistic material), the latter two of which seemed especially relevant to the linguistic target and types of treatment activities in the present experiment. It takes roughly a half hour to administer, using an audio CD to control the timing of the various sections.

In Part III, participants are given 5 minutes to answer 50 items in which a quasi-phonetically spelled word is followed by 5 options and their task is to select the one that “corresponds most nearly in meaning with the disguised word” (MLAT test booklet, p. 2). An example provided as part of the instructions is the word ernst, which is followed by the following choices: (A) shelter, (B) sincere, (C) slanted, (D) free, and (E) impatient. The intended answer is given as (B) since “ernst is a disguised spelling of earnest, which corresponds most nearly in meaning to sincere” (p. 2).

In Part IV, participants are given 15 minutes to answer 45 items thought to measure their grammatical sensitivity. After reading a “key” sentence in which one word is underlined and printed in capital letters, they must read another sentence or group of sentences in which 5 words and/or phrases are underlined and marked A, B, C, D, and E. The participants’ task is to choose the underlined word or phrase in this second (set of) sentence(s) which “does the same thing” as the capitalized word in the key sentence (p. 5). An example from the instructions reads as follows:

Mary is cutting the APPLE.

My brother John is beating his dog with a big stick.

A  B  C  D  E
It is explained that since “APPLE is the name of the thing which is being cut” in the key sentence, and “dog is the thing which is being beaten” in the second sentence, the answer is D (p. 5).

In Part V, participants first study 24 written Kurdish-English translation equivalents (including nouns, verbs, adjectives, and prepositions) for 2 minutes. They then practice recalling the English versions for another 2 minutes by filling in blanks provided next to Kurdish words on a practice exercise sheet. Finally, they are given 4 minutes to answer 24 multiple-choice items in which a Kurdish word is followed by 5 choices of possible equivalents in English. All parts of the MLAT Short Form are scored by awarding 1 point for each correct response and 0 points for an incorrect response. A scoring key is provided along with the test materials.

The Visual Patterns Test

A variety of tests exist for the measurement of visual memory. Some, such as the Corsi Blocks Test (Corsi, 1972) and the Visual Span subtest of the revised Wechsler Memory Scale (Wechsler, 1987), involve remembering the temporal dynamics of a sequence of spatial locations, whereas others arguably tap a more purely visual capacity. For the present study, it was decided to use Della Sala et al.’s (1997) Visual Patterns Test (VPT), published by Harcourt Assessment for the Thames Valley Test Company. The VPT has been shown to disentangle visual storage capacity from other aspects of short-term memory, including temporal sequencing (Della Sala et al., 1999). Participants are shown checkerboard-like grids of blank and filled-in squares which are designed to be virtually impossible to code in verbal form. Two sample items (the very first and the very last from Version A of the VPT) are shown in Figure 4.9.
During the test, which is designed to be administered individually by a researcher, participants are presented with sets of matrices in which half of the boxes are filled in. Their size gradually increases from 2 x 2 through 5 x 6 over the course of the test, adding 2 boxes for each level increase. There are 14 levels, with 3 items of the same size and shape at each level, for a total of 42 items. Thus, the first 3 items involve grids made up of 4 boxes with 2 filled in; the next 3 items involve grids made up of 6 boxes with 3 filled in, and so on, until the last 3 grids are made up of 30 boxes with 15 filled in.

The test materials consist of a set of stimulus cards, which the tester presents to the participant one at a time, and a set of answer sheets (an example of which is provided in Appendix J). Each answer sheet is organized in two columns, one for each level, with the 3 blank answer grids for each level arranged vertically. The dimensions of the blank answer grids correspond exactly in size and shape to the partially filled-in grids on the tester’s stimulus cards. In order to ensure that answers from previous items do not interfere with the current item, the tester folds each answer sheet longitudinally before giving it to the participant. The test manual states that each card should be shown to the participant for 3 seconds and then covered up.
while the participant responds by marking the appropriate boxes on the blank answer grid corresponding to that item. There is no time limit for responding, and participants are allowed to modify their answers as much as they would like before presenting a response to the tester. At that point, if the number of boxes the participant has filled in does not match the number in the target pattern, according to the manual, “the subject should be encouraged to add or delete filled squares until the numbers agree” (p. 8). Finally, the tester reveals the stimulus card, providing feedback on each item.

There are 2 parallel versions of the VPT: Version A and Version B. For the present experiment, only Version A was used unless something happened to disrupt the administration of an item (e.g., if another student suddenly entered the room), in which case an equivalent item from Version B was substituted. This occurred very rarely. One difference from the administration procedures described above was in the amount of time each stimulus card was presented to participants. Piloting had shown it to be unfeasible to use a stopwatch or clock to measure the 3 seconds of presentation time precisely while also managing the stack of stimulus cards; thus, I kept time by mentally counting, “one, one thousand; two, one thousand; three, one thousand.” This actually amounted to 4-5 seconds as opposed to 3. Another modification of the administration involved the feedback that was provided. Instead of repeatedly encouraging participants to add or delete squares until the right number was provided, I simply gave feedback one time per item, indicating how many squares needed to be added, deleted, and/or moved. The instructions, which I read to the participants directly from the test booklet (p. 8), were as follows:
This is a test of visual memory for patterns. You will be shown a pattern like this [indicate the example on the card] and asked to recall it by making marks in these blank squares [show the subject the appropriate blank grid]. You will only be shown the pattern for a short time, so you will have to concentrate hard in order to remember it. As soon as I cover up the pattern, you may start to respond.

The patterns are simple to start with and become more complicated later.

For the purposes of scoring, the manual states that a participant’s “raw visual span” corresponds to the highest level of grid complexity at which at least 1 of the 3 patterns is answered completely accurately. However, it also notes that “[f]or scientific research purposes, a more sensitive measure of recall would be provided by the mean of the complexities of, say, the last three patterns recalled” (p. 8). In the current context, to include even more information in the scoring, points were awarded for every item that a participant answered correctly, and each item was worth the number of squares that were filled in. In other words, on a 3 x 4 grid with 6 boxes filled in, a participant could score 6 points for a completely correct answer. However, if only 5/6 of the participant’s filled-in boxes were in the appropriate positions, 0 points would be awarded. The total number of points was then divided by the number of items (42). Thus, each participant’s final score represents an average of the number of filled-in squares involved when the participant got an item completely right. For the 79 participants who completed the VPT, the mean was 5.81, with a standard deviation of 1.15 and a range from 2.52 to 8.17.
Summary of variables

The research questions outlined at the beginning of this chapter centered around three issues: (1) whether providing learners with tree diagrams as feedback on their interpretations of L2 sentences with reflexives would enable them to interpret such sentences more accurately; (2) whether any of a variety of individual differences would be related to the learners’ accuracy in interpreting L2 sentences with reflexives; and (3) whether the relationships among these individual differences and the learners’ accuracy of interpretation would differ according to treatment condition. In view of the large number of variables involved, it will be helpful to review them here.

Relevant to the first research question are the independent variable of treatment group (No Feedback, Right/Wrong Feedback, or Trees Feedback) and the dependent variables of the participants’ scores on the pre-, post-, and delayed post-tests of zibun interpretation. For the second research question, relevant variables include the participants’ zibun test scores, their scores on several individual-difference measures, and the characteristics reported on the background questionnaire. The IDs to be focused on here include the participants’ visual short-term memory (as measured by the Visual Patterns Test), grammatical sensitivity (Modern Language Aptitude Test, Part IV), rote memory for linguistic material (MLAT, Part V), sensitivity to ambiguity in English sentences (including, more specifically, sensitivity to ambiguity in English sentences containing reflexives), knowledge of metalinguistic terminology, reported enjoyment of grammar, years of Japanese study, Japanese course level, and number of linguistics courses taken. For the third research question, all of these variables are relevant.
Chapter 5 presents the results of the analyses which have been conducted so far to answer the research questions. Data left for future research include the confidence judgments and source attributions which the participants made on each test item, their scores on items theoretically involving positive versus negative evidence, the accuracy of their sentence interpretations during the treatment sessions, their working memory capacities (as measured by listening-span and operation-span tests), their levels and types of motivation over the course of the study (assessed via pre-actional, actional, and post-actional motivation questionnaires), their responses on the reflection questionnaire, and the introspective data from the concurrent verbalizations which half of the participants produced during the treatment sessions. Some possibilities for further analysis are discussed in Chapter 6.
CHAPTER 5

RESULTS

Introduction

This chapter presents the results of the statistical analyses which were carried out in order to answer the research questions. The analyses were conducted using PASW (Predictive Analytics SoftWare) Statistics 18, from SPSS: An IBM Company, and the alpha level was set at .05 for all tests. The results are organized by research question. Each section will begin by discussing whether the assumptions for parametric statistics (e.g., normality of distributions, homogeneity of variances, sphericity, etc.) were met. When the assumptions of a statistical test are violated, according to Larson-Hall (2010, citing Wilcox, 2001), this “does not mean your results are invalid” (p. 355). What it does mean, however, is that there may be less power to find relationships that do exist (i.e., a Type II error). Thus, in some cases, where possible, nonparametric statistics were chosen as more appropriate alternatives. Next, descriptive statistics (means, ranges, and standard deviations) will be presented for each dependent and predictor variable, both overall (i.e., including all participants) and broken down by experimental condition. Since different experimental groups are compared, it will be verified that they did not show statistically significant differences from each other on certain variables of interest at the outset of the study (e.g., pre-test scores, cognitive measures). Once this has been done, the results of the inferential statistics used to answer the research question at hand will be presented.
Research questions involving the effectiveness of feedback

The first set of research questions asked whether L2 learners of Japanese would show improvement in their ability to understand what sentences containing the reflexive zibun can and cannot mean following one of three conditions of interpretation practice: (1) without feedback (No Feedback), (2) with feedback indicating whether each interpretation was right or wrong (Right/Wrong Feedback), or (3) with feedback indicating whether each interpretation was right or wrong, followed by a tree diagram illustrating the underlying structure of the sentence (Trees Feedback). Centrally of interest also was whether the amounts of improvement shown, if any, would be statistically different among the three conditions. In order to answer these questions, a series of repeated-measures analyses of variance (RM ANOVAs) were run, first to examine the trajectories of pre-, post-, and delayed post-test scores within each experimental group independently, and then to compare the groups’ improvements (or lack thereof) against one another.

Assumptions of repeated-measures analyses of variance

Repeated-measures ANOVAs are appropriate to use when the same participants have been tested more than once and more than two means are being compared. They carry the assumptions of normality of distributions, homogeneity of variances, and sphericity. Table 5.1 displays several statistics relevant to checking for normality of distributions: skewness and kurtosis values, along with the results of Kolmogorov-Smirnov and Shapiro-Wilk goodness-of-fit tests. Noting that such formal tests often do not have enough power to detect violations of normality, Larson-Hall (2010) argues (citing Wilcox, 2003, and Wilkinson & the Task Force on...
Statistical Inference, 1999) that it is crucial to examine the descriptive statistics and visual
displays of the data as well (pp. 84-5). Thus, Figure 5.1 and Table 5.2 provide additional
perspectives on the distributional characteristics of the data. Figure 5.1 shows boxplots of the
groups’ test scores, and Table 5.2 displays the means, ranges, standard deviations, variances,
and 95% confidence intervals for each group on each test.

Figure 5.1. Boxplots of test scores across treatment conditions
Table 5.1. Goodness-of-fit tests to check for normality of distributions on the pre-, post-, and delayed post-tests across groups

<table>
<thead>
<tr>
<th>Test</th>
<th>Treatment</th>
<th>Skewness</th>
<th></th>
<th>Kurtosis</th>
<th></th>
<th>Kolmogorov-Smirnov$^\S$</th>
<th>Shapiro-Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Skew</td>
<td>Std. Error</td>
<td>Ratio</td>
<td>Kurtosis</td>
<td>Std. Error</td>
<td>Ratio</td>
</tr>
<tr>
<td>Pre-test</td>
<td>No Feedback</td>
<td>-0.47</td>
<td>0.46</td>
<td>1.01</td>
<td>0.08</td>
<td>0.90</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Right/Wrong Feedback</td>
<td>-0.06</td>
<td>0.43</td>
<td>0.13</td>
<td>-0.12</td>
<td>0.85</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Trees Feedback</td>
<td>-0.58</td>
<td>0.47</td>
<td>1.24</td>
<td>-0.65</td>
<td>0.92</td>
<td>0.71</td>
</tr>
<tr>
<td>Post-test</td>
<td>No Feedback</td>
<td>-0.10</td>
<td>0.46</td>
<td>0.22</td>
<td>-0.86</td>
<td>0.90</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Right/Wrong Feedback</td>
<td>0.31</td>
<td>0.43</td>
<td>0.71</td>
<td>-0.75</td>
<td>0.85</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>Trees Feedback</td>
<td>0.09</td>
<td>0.47</td>
<td>0.19</td>
<td>-0.54</td>
<td>0.92</td>
<td>0.59</td>
</tr>
<tr>
<td>Delayed post-test</td>
<td>No Feedback</td>
<td>0.63</td>
<td>0.46</td>
<td>1.35</td>
<td>0.03</td>
<td>0.90</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Right/Wrong Feedback</td>
<td>0.60</td>
<td>0.43</td>
<td>1.38</td>
<td>-0.27</td>
<td>0.85</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>Trees Feedback</td>
<td>0.32</td>
<td>0.47</td>
<td>0.68</td>
<td>-0.07</td>
<td>0.92</td>
<td>0.08</td>
</tr>
</tbody>
</table>

$\S$Performed with the Lilliefors correction
Table 5.2. Descriptive statistics of pre-, post-, and delayed post-test scores across groups

<table>
<thead>
<tr>
<th>Test</th>
<th>Treatment Condition</th>
<th>Mean</th>
<th>Range</th>
<th>Std. Dev.</th>
<th>Variance</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Pre-test</td>
<td>No Feedback</td>
<td>58.76</td>
<td>40.52</td>
<td>10.17</td>
<td>103.40</td>
<td>54.56</td>
</tr>
<tr>
<td></td>
<td>Right/Wrong Feedback</td>
<td>59.39</td>
<td>38.57</td>
<td>9.51</td>
<td>90.48</td>
<td>55.77</td>
</tr>
<tr>
<td></td>
<td>Trees Feedback</td>
<td>59.47</td>
<td>46.43</td>
<td>13.52</td>
<td>182.91</td>
<td>53.76</td>
</tr>
<tr>
<td>Post-test</td>
<td>No Feedback</td>
<td>64.72</td>
<td>48.55</td>
<td>13.56</td>
<td>183.77</td>
<td>59.12</td>
</tr>
<tr>
<td></td>
<td>Right/Wrong Feedback</td>
<td>65.09</td>
<td>49.53</td>
<td>13.40</td>
<td>179.47</td>
<td>60.00</td>
</tr>
<tr>
<td></td>
<td>Trees Feedback</td>
<td>73.77</td>
<td>53.33</td>
<td>14.48</td>
<td>209.55</td>
<td>67.66</td>
</tr>
<tr>
<td>Delayed post-test</td>
<td>No Feedback</td>
<td>62.16</td>
<td>45.96</td>
<td>11.62</td>
<td>135.00</td>
<td>57.37</td>
</tr>
<tr>
<td></td>
<td>Right/Wrong Feedback</td>
<td>64.20</td>
<td>46.43</td>
<td>12.17</td>
<td>148.07</td>
<td>59.57</td>
</tr>
<tr>
<td></td>
<td>Trees Feedback</td>
<td>70.74</td>
<td>53.33</td>
<td>12.69</td>
<td>161.07</td>
<td>65.38</td>
</tr>
</tbody>
</table>

No Feedback n=25, Right/Wrong Feedback n=29, Trees Feedback n=24
Skewness refers to a deviation in the shape of a distribution with respect to the symmetry of scores around the mean. The skewness ratio is the skewness level divided by the standard error of the skewness. According to Weinberg and Abramowitz (2002, p. 278), a ratio greater than 2 indicates a departure from normality. As shown in Table 5.1, all skewness ratios are between 0.13 and 1.38; thus, they do not provide evidence that the assumption of normality was violated. Kurtosis refers to a deviation in the shape of a distribution with respect to the concentration of scores around the mean. All of the kurtosis ratios displayed in Table 5.1 fall between 0.03 and 0.95, likewise no cause for concern. For the Kolmogorov-Smirnov and Shapiro-Wilk tests, $p$-values less than 0.05 would indicate a statistically significant difference from a normal distribution. All of the $p$-values in Table 5.1 are above 0.05, again providing no evidence to reject the hypothesis of normal distributions.

The assumption of homogeneity of variances is that the variances (or the standard deviations) of the data should be similar across the experimental groups of a study. Levene’s test checks the null hypothesis of equal variances; thus, as above, if $p$-values are greater than 0.05, we do not have strong evidence to believe that the variances are different from each other (Larson-Hall, 2010, p. 88). As shown in Table 5.3, the $p$-values of Levene’s tests for the pre-, post-, and delayed post-tests are all greater than 0.05, meaning that we can provisionally consider the assumption of homogeneity of variances to have been met—although, again, Larson-Hall (2010) notes that, given controversies in the statistics literature regarding whether such tests are powerful enough, “[m]ore informal ways of testing equal variances are probably better, such as looking at boxplots and numerical summaries” (p. 278). This will be done below.
The last assumption, sphericity, is essentially an assessment of the homogeneity of variances for the same person across repeated measures—“whether the differences between the variances of an individual participant’s data are equal” (Larson-Hall, 2010, p. 401). It can be measured by Mauchly’s test of sphericity. In this case, the results in Table 5.3 indicate that the assumption has been violated; the p-value is less than 0.01, meaning that we should reject the assumption that sphericity holds. Howell (2002) recommends some possible approaches to take in this situation: (1) to use a Greenhouse-Geisser correction, available in the PASW output, for the group and interaction effects (in fact, he recommends this regardless of whether the sphericity assumption is met), and (2) to conduct a separate repeated-measures ANOVA for each experimental group—attempting, however, to control for the increased possibility of Type I errors by deciding beforehand which measures are most important to compare and focusing only on those (Larson-Hall, 2010, p. 337).

Table 5.3. Tests of homogeneity and sphericity assumptions for Research Question 1

<table>
<thead>
<tr>
<th>Test</th>
<th>Levene’s Test of Homogeneity of Variances</th>
<th>Mauchly’s Test of Sphericity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>df1</td>
</tr>
<tr>
<td>Pre-test</td>
<td>2.060</td>
<td>2</td>
</tr>
<tr>
<td>Post-test</td>
<td>0.086</td>
<td>2</td>
</tr>
<tr>
<td>Delayed post-test</td>
<td>0.048</td>
<td>2</td>
</tr>
</tbody>
</table>
Finally, we turn back to Figure 5.1 for a visual inspection of the data’s distributional characteristics. The whisker lines in the boxplots represent the ranges (excluding outliers), shaded boxes represent the interquartile ranges (IQRs, from the 25th to 75th percentiles), and horizontal lines in the middles of the boxes represent the medians. If a distribution is normal (as opposed to skewed), the median line will be in the middle of its box. If two distributions have equal variances, the lengths of the shaded boxes (IQRs) will be similar. Outliers (of which there are none here) would be represented with labeled circles for any data points falling outside 1.5 times the IQR. Looking at the test-score distributions across groups, although the Trees group might show slightly more variance than the other two, the differences do not appear to be extreme, nor do the data seem to be strongly skewed.

In sum, the data look to be sufficiently normal in distribution, with sufficient homogeneity of variances, for ANOVAs to be run. Since the sphericity assumption was not met, the approach, following Howell’s (2002) suggestions, was (1) to perform separate RM ANOVAs in order to examine each group’s improvement over time independently (already planned as part of the first set of research questions), and (2) to use a Greenhouse-Geisser correction in a repeated-measures ANOVA comparing the amounts of improvement shown by the groups relative to each other. Type II Sums of Squares (SS) were used in each case.34

34 Most statistical programs, including PASW, use Type III SS by default. In practice, this choice matters only for main effects and only when group sizes differ, but see Larson-Hall (2010, pp. 311-313, citing Fox, 2002; Langsrud, 2003; Nelder, 1994; and Nelder & Lane, 1995) for arguments that Type II SS, which assume that interaction effects are not statistically significant, are more logical, have more power, and test more realistic hypotheses.
Effects of the experimental manipulation

Figure 5.2 displays the three treatment groups’ average scores on the pre-test, post-test, and delayed post-test. All three groups’ mean pre-test scores are within 1 percentage point of each other (between 58.76 to 59.47%), and, according to a one-way ANOVA, the differences among them are not statistically significant ($F(2,75) = 0.031, p=0.97$). Thus, the groups can be considered comparable with regard to their test scores at the outset. On the post-tests, there are larger differences in the means, with the Trees-Feedback group appearing to score higher than the other two on both the post-test (No-Feedback: 64.72%, Right/Wrong: 65.09%, Trees: 73.77%) and the delayed post-test (No-Feedback: 62.16%, Right/Wrong: 64.20%, Trees: 70.74%). Whereas scores in the Trees group increased, on average, roughly 14% from pre- to post-test and 11% from pre- to delayed post-test, scores in the No-Feedback and Right/Wrong-Feedback groups increased only between 3 to 6% across these tests. It is also worth noting that, on both of the post-tests, the 95% confidence intervals of the No-Feedback and Right/Wrong-Feedback groups are largely overlapping, whereas they overlap much less with the 95% confidence interval for the Trees-Feedback group. For a finer-grained perspective on the data, Appendix M displays parallel coordinate plots, one for each group, with a separate line showing the performance of each learner in that condition.

To investigate whether the learners in each treatment condition showed statistically significant improvement over the course of the tests, a RM ANOVA was run for each group separately, with Time (Pre, Post, Delay) as the within-subjects variable. These results are displayed in Table 5.4. A group can be considered to have shown improvement if the effect for Time is statistically significant at $p < 0.05$. The column entitled Partial Eta$^2$ (reported in the text
as $\eta^2_p$) indicates the magnitude of each effect. According to Brown (2008, pp. 40, 42), $\eta^2_p$ can be defined as the percentage of variance in an effect (or interaction) and its associated error which is accounted for by that effect (or interaction). It is used with ANOVAs when the data are not independent (i.e., when the same people contribute to more than one cell, as in the repeated-measures design here). Cohen (1988) cautions that the meanings of effect sizes are relative to the area of research and the experimental methods employed, but one possible rule of thumb for interpreting $\eta^2$ values (based on a combination of guidelines from Cohen (1992) and a conversion table he provides in his 1988 book) would be that effect sizes of 0.01, 0.06, and 0.14 might be considered small, medium, and large, respectively.

Table 5.4. Separate RM ANOVAs testing for improvement over time in each treatment group

<table>
<thead>
<tr>
<th>Tests of Within-Subjects Effects</th>
<th>Type II SS</th>
<th>df</th>
<th>F</th>
<th>p-value</th>
<th>Partial Eta$^2$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No-Feedback Condition (n=25)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>446.330</td>
<td>1.874</td>
<td>2.902</td>
<td>.069</td>
<td>.108</td>
<td>.522</td>
</tr>
<tr>
<td>Error(Time)</td>
<td>3691.641</td>
<td>44.978</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Right/Wrong-Feedback Condition (n=29)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>545.972</td>
<td>1.767</td>
<td>2.330</td>
<td>.114</td>
<td>.077</td>
<td>.423</td>
</tr>
<tr>
<td>Error(Time)</td>
<td>6559.792</td>
<td>49.479</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Trees-Feedback Condition (n=24)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>2723.744</td>
<td>1.653</td>
<td>12.417</td>
<td>.000</td>
<td>.351</td>
<td>.994</td>
</tr>
<tr>
<td>Error(Time)</td>
<td>5045.013</td>
<td>38.030</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Performed with Greenhouse-Geisser corrections
For the No-Feedback condition, in which participants practiced interpreting sentences without receiving feedback, the effect for Time is not quite statistically significant ($F(1.87, 44.98) = 2.902, p = 0.07, \eta^2_p = 0.11$), meaning that we do not have evidence that would allow us to reject the null hypothesis of no improvement over the course of all three tests. When paired $t$-tests\textsuperscript{35} are used to perform pairwise comparisons between only two of the tests at a time, however, it appears that although the No-Feedback group did not show statistically significant improvement from the pre-test to the delayed post-test ($t(24) = -1.353, p = 0.19, d = 0.31$), they did demonstrate improvement from the pre-test to the immediate post-test ($t(24) = -2.184^*, p$\textsuperscript{*}).

\textsuperscript{35} A paired $t$-test is equivalent to a RM ANOVA with fewer than 3 measures being compared. The PASW output does not provide effect sizes for $t$-tests; thus, an online calculator (http://web.uccs.edu/lbecker/Psy590/escalc3.htm) was used to calculate Cohen’s $d$. 

\textbf{Figure 5.2.} Comparison of the treatment conditions with regard to mean test scores

![Chart](Chart.png)
Cohen (1988) suggests that when using $d$ as a measure of effect size, values of 0.2, 0.5, and 0.8 can be considered small, medium, and large, respectively. Thus, there may be some evidence of a medium-sized learning effect in the No-Feedback condition, albeit possibly short-lived.

For the Right/Wrong-Feedback condition, in which participants practiced interpreting sentences and received messages informing them of whether their interpretations had been right or wrong, the effect for Time is not statistically significant ($F(1.77, 49.48) = 2.330, p = 0.11, \eta^2_p = 0.08$). In this condition, pairwise comparisons between pre- and post-test scores ($t(28) = -1.786, p = 0.09, d = 0.49$) and pre- and delayed post-test scores ($t(28) = -1.625, p = 0.12, d = 0.44$) are not statistically significant, either; thus, we do not have evidence which would allow us to reject the null hypothesis of no improvement.

For the Trees-Feedback condition, in which the participants practiced interpreting sentences and received both right/wrong messages and tree diagrams illustrating the underlying structures of the sentences, the effect for Time is larger in magnitude and statistically significant ($F(1.65, 38.03) = 12.417^*, p <0.001, \eta^2_p = 0.35$), allowing us to reject the null hypothesis of no improvement. In this case, paired t-tests indicate statistically significant differences in test scores, with large effect sizes, both from pre-test to immediate post-test ($t(24) = -4.268^*, p <0.001, d = 1.02$) and from pre-test to delayed post-test ($t(24) = -3.006^*, p = 0.006, d = 0.86$). This can be taken as evidence of learning that persisted at least a week beyond the last time feedback had been provided.

It should be noted that the power to find a significant effect for Time in the RM ANOVAs was quite low for the No-Feedback and Right/Wrong-Feedback conditions (0.52 and 0.42,
respectively). A power of 0.5 means only a 50% chance of finding an effect if one exists. The consensus within the statistics community, according to Murphy and Myors (2004), is that researchers can consider a power level of 0.8 to be sufficient. Thus, the level of power in the Trees condition (0.99) was more than adequate for finding an effect, whereas in the other conditions it was not. Collecting data from a greater number of learners might have produced statistically significant results for the No-Feedback and Right/Wrong-Feedback groups, but the amount of improvement demonstrated would likely not have been great in practical terms. Considering the smaller effect sizes in those groups, it is not surprising that the power was low.

Finally, to ascertain whether there were statistical differences among the treatment groups with respect to how much improvement they demonstrated (or did not demonstrate), a RM ANOVA was run including all participants, with Time (Pre, Post, and Delay) as the within-subjects variable and Treatment (No Feedback, Right/Wrong Feedback, and Trees Feedback) as the between-subjects variable. In this case, an affirmative answer to the question would require a statistically significant Time-by-Treatment interaction effect, indicating that the groups showed different trajectories in their test performance over time. The results of the repeated-measures ANOVA are presented in Table 5.5.

Looking first at the within-subjects effect for Time ($F(1.78, 133.58) = 14.742^*, p < 0.001, \eta_p^2 = 0.16$), the statistically significant $F$-value allows us to reject the null hypothesis of no improvement in general; that is, considered together regardless of treatment condition, the participants showed improved performance over the course of the tests. The Time-by-Treatment interaction effect, however, is small and not statistically significant ($F(3.56, 133.58) = 1.739, p = 0.15, \eta_p^2 = 0.04$). In other words, despite the fact that we previously found evidence
for sustained improvement in the Trees condition on its own, while not finding such strong
evidence for the No-Feedback and Right/Wrong conditions on their own, it is not possible to
say based only on the present result that the three experimental groups differed from one
another in their trajectories of improvement. The power level of 0.49 suggests that the
probability of finding such a difference if one existed was unacceptably low, although it makes
sense in light of the small effect size. Last but not least, the between-subjects effect for
Treatment closely approaches statistical significance ($F(2, 75) = 3.091^0, p = 0.051, \eta_p^2 = 0.08$),
indicating a trend toward the groups performing differently from each other overall, ignoring
time. Here again, the observed power to find a true difference was insufficient (a 58% chance).

Table 5.5. Results of RM ANOVA comparing test scores in the 3 treatment groups over time

<table>
<thead>
<tr>
<th>Source</th>
<th>Type II Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>$F$</th>
<th>$p$-value</th>
<th>Partial Eta$^2$</th>
<th>Obs. Power</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tests of Within-Subjects Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>3006.59</td>
<td>1.78</td>
<td>1688.03</td>
<td>14.742</td>
<td>.000</td>
<td>.164</td>
<td>.998</td>
</tr>
<tr>
<td>Time x Treatment</td>
<td>709.46</td>
<td>3.56</td>
<td>199.16</td>
<td>1.739</td>
<td>.152</td>
<td>.044</td>
<td>.489</td>
</tr>
<tr>
<td>Error(Time)</td>
<td>15296.45</td>
<td>133.58</td>
<td>114.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Test of Between-Subjects Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>962641.34</td>
<td>1</td>
<td>962641.34</td>
<td>3746.39</td>
<td>.000</td>
<td>.980</td>
<td>1.000</td>
</tr>
<tr>
<td>Treatment</td>
<td>1588.25</td>
<td>2</td>
<td>794.13</td>
<td>3.091</td>
<td>.051</td>
<td>.076</td>
<td>.579</td>
</tr>
<tr>
<td>Error</td>
<td>19271.38</td>
<td>75</td>
<td>256.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Performed with Greenhouse-Geisser corrections
Research questions involving aptitude and aptitude-treatment interactions

The second set of research questions asked whether individual differences (IDs) were related to learners’ test performance and to the amount of improvement they showed over the course of the tests. Among the IDs investigated were grammatical sensitivity and rote memory for linguistic material (aspects of L2 aptitude measured by the MLAT), visual short-term memory (STM), knowledge of metalinguistic terminology, sensitivity to ambiguity in English (including for sentences containing reflexive pronouns specifically), years of Japanese study, Japanese course level, number of linguistics courses taken, and enjoyment of grammar. In order to answer these questions, correlational analyses were conducted: bivariate correlations between ID measures on the one hand and test scores and gain scores on the other, and partial correlations between ID measures and post-test scores, controlling for pre-test scores.

The third set of research questions, closely related to those in the second set, involved the issue of aptitude-treatment interactions; in other words, they focused on whether the relationships between ID variables and test scores differed according to treatment condition. In answering them, the same statistical tests were employed, the difference being simply that analyses were conducted separately for each group and then compared across groups. Thus, the next section will address whether the assumptions for parametric statistics were met both overall (i.e., including all participants) and separately by group (including only the participants in each treatment condition).36

36 Multiple regression analyses were also conducted to explore which combinations of these ID variables could independently explain the most variance in test scores, both overall and within each experimental group; however, as will be discussed below, due to insufficient sample size, these analyses did not seem to add information beyond what had already been gathered from the canonical correlations.
Assumptions underlying the use of correlations

Bivariate correlations are used to investigate the extent to which two continuous variables vary together. Assumptions underlying the use of parametric correlations include that the data have been gathered independently (i.e., that the scores of one person do not influence the scores of another person), that each of the variables is normally distributed, that the relationship between the variables is linear, and that the variances of their residuals are similar (known as homoscedasticity). Checking the assumptions of normality of distributions, linearity, and homoscedasticity involves examining boxplots and histograms, scatterplots with Loess and regression lines, and plots of studentized residuals against fitted values, respectively, as explained in more detail below.

The distributions of the participants’ pre-, post-, and delayed post-test scores were determined to be sufficiently normal in the previous section; thus, this section will focus on the distributions of the ID variables. Table 5.6 displays descriptive statistics (means and standard deviations) for each ID measure, both overall and by experimental group. The last column in this table presents the results of Levene’s test of homogeneity of variances.37 Table 5.7 shows the skewness and kurtosis ratios for each ID measure, both overall and by experimental group. Again, if these are outside the range of -2 to 2, the data are considered to be skewed or leptokurtic. Histograms, boxplots, and/or scatterplots are included in Appendices O1-O9 for visual inspection of the data’s distributional characteristics.

---

37 Homogeneity of variances is not strictly an assumption underlying the use of correlations, but it may be relevant here for two reasons: (1) one-way ANOVAs, for which homogeneity is an assumption, will be used to check for differences among the groups on the various ID measures, and (2) correlations between IDs and test scores will later be run for each group separately and then compared across groups. If one group were to show a large amount of variance on an ID measure, while the scores of another group were more compressed, it might be easier to find a co-varying relationship in the former group. That could be important in interpreting the results.
Table 5.6. Descriptive statistics and tests of homogeneity for each ID measure

<table>
<thead>
<tr>
<th>ID Measure</th>
<th>No Feedback (n=25 except Enjoy Grammar [24])</th>
<th>Right/Wrong (n=29 except VPT, Enjoy Grammar [28])</th>
<th>Trees (n=26)</th>
<th>All (N=80 except VPT [79], Enjoy Grammar [78])</th>
<th>Levene Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPT</td>
<td>5.41&lt;sup&gt;a&lt;/sup&gt; (0.95)</td>
<td>6.18&lt;sup&gt;a&lt;/sup&gt; (1.13)</td>
<td>5.79 (1.25)</td>
<td>5.81 (1.15)</td>
<td>.504</td>
</tr>
<tr>
<td>MLAT IV</td>
<td>25.72 (5.82)</td>
<td>25.93 (5.97)</td>
<td>27.54 (6.36)</td>
<td>26.39 (6.03)</td>
<td>.052</td>
</tr>
<tr>
<td>MLAT V</td>
<td>20.00 (4.61)</td>
<td>21.55 (2.60)</td>
<td>21.35 (3.75)</td>
<td>21.00 (3.71)</td>
<td>3.300*</td>
</tr>
<tr>
<td>MLAT Total</td>
<td>69.64 (14.27)</td>
<td>72.03 (13.08)</td>
<td>77.15 (14.61)</td>
<td>72.95 (14.13)</td>
<td>.039</td>
</tr>
<tr>
<td>Reflexives Subtotal</td>
<td>6.46 (2.05)</td>
<td>6.55 (2.20)</td>
<td>7.14 (1.87)</td>
<td>6.71 (2.05)</td>
<td>.789</td>
</tr>
<tr>
<td>Ambiguity Subtotal</td>
<td>11.88 (2.86)</td>
<td>12.35 (3.12)</td>
<td>13.50 (2.71)</td>
<td>12.58 (2.95)</td>
<td>.642</td>
</tr>
<tr>
<td>Grammar Subtotal</td>
<td>6.46 (3.34)</td>
<td>6.55 (3.24)</td>
<td>6.10 (3.83)</td>
<td>6.38 (3.43)</td>
<td>1.406</td>
</tr>
<tr>
<td>Ambiguity &amp; Grammar Total</td>
<td>18.34 (4.76)</td>
<td>18.90 (4.81)</td>
<td>19.60 (5.87)</td>
<td>18.95 (5.12)</td>
<td>1.101</td>
</tr>
<tr>
<td>Years of Japanese Study</td>
<td>4.70 (2.27)</td>
<td>3.64 (1.53)</td>
<td>4.62 (2.53)</td>
<td>4.29 (2.16)</td>
<td>4.311*</td>
</tr>
<tr>
<td>Course Level</td>
<td>3.36 (1.00)</td>
<td>3.66 (1.50)</td>
<td>3.96 (1.66)</td>
<td>3.66 (1.42)</td>
<td>2.067</td>
</tr>
<tr>
<td># Linguistics Courses</td>
<td>1.48 (2.22)</td>
<td>1.38 (2.57)</td>
<td>1.46 (2.12)</td>
<td>1.44 (2.29)</td>
<td>.306</td>
</tr>
<tr>
<td>Enjoyment of Grammar</td>
<td>3.42 (1.14)</td>
<td>3.46&lt;sup&gt;b&lt;/sup&gt; (1.26)</td>
<td>2.69&lt;sup&gt;b&lt;/sup&gt; (1.52)</td>
<td>3.19 (1.35)</td>
<td>1.101</td>
</tr>
</tbody>
</table>

Notes:

Means are presented with the standard deviations in parentheses.

Statistically significant results for Levene’s tests, indicated with asterisks, suggest violations of the assumption of homogeneity of variances.

Superscripts (a, b) indicate where post-hoc tests identified differences in the means.
Table 5.7. Skewness and kurtosis ratios for each ID, both overall and by treatment group

<table>
<thead>
<tr>
<th>ID Measure</th>
<th>No Feedback</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Skew</td>
<td>Kurtosis</td>
<td>Skew</td>
<td>Kurtosis</td>
<td>Skew</td>
<td>Kurtosis</td>
<td>Skew</td>
<td>Kurtosis</td>
</tr>
<tr>
<td>VPT</td>
<td>-1.138</td>
<td>-0.703</td>
<td>-1.673</td>
<td>1.150</td>
<td>-0.976</td>
<td>0.598</td>
<td>-1.502</td>
<td>0.290</td>
</tr>
<tr>
<td>MLAT IV</td>
<td>0.778</td>
<td>0.147</td>
<td>1.037</td>
<td>-0.813</td>
<td>-1.844</td>
<td>1.359</td>
<td>-0.089</td>
<td>-0.421</td>
</tr>
<tr>
<td>MLAT V</td>
<td>-2.254</td>
<td>-0.361</td>
<td>-1.581</td>
<td>-1.084</td>
<td>-2.910</td>
<td>0.448</td>
<td>-4.781</td>
<td>1.333</td>
</tr>
<tr>
<td>MLAT Total</td>
<td>0.567</td>
<td>0.014</td>
<td>0.182</td>
<td>-1.614</td>
<td>-2.048</td>
<td>0.393</td>
<td>-0.669</td>
<td>-1.430</td>
</tr>
<tr>
<td>Reflexives Subtotal</td>
<td>-1.106</td>
<td>-0.981</td>
<td>-0.972</td>
<td>-1.388</td>
<td>-1.171</td>
<td>-0.986</td>
<td>-1.825</td>
<td>-1.786</td>
</tr>
<tr>
<td>Ambiguity Subtotal</td>
<td>-1.856</td>
<td>0.538</td>
<td>-0.555</td>
<td>-0.968</td>
<td>-0.226</td>
<td>-1.444</td>
<td>-1.420</td>
<td>-0.654</td>
</tr>
<tr>
<td>Grammar Subtotal</td>
<td>1.149</td>
<td>-0.213</td>
<td>0.592</td>
<td>0.156</td>
<td>1.020</td>
<td>-0.852</td>
<td>1.413</td>
<td>-0.835</td>
</tr>
<tr>
<td>Ambiguity &amp; Grammar Total</td>
<td>-0.058</td>
<td>-0.600</td>
<td>-0.157</td>
<td>-0.895</td>
<td>0.261</td>
<td>-0.831</td>
<td>0.312</td>
<td>-1.162</td>
</tr>
<tr>
<td>Years of Japanese Study</td>
<td>1.858</td>
<td>0.908</td>
<td>1.634</td>
<td>-0.021</td>
<td>1.309</td>
<td>-0.871</td>
<td>3.309</td>
<td>0.652</td>
</tr>
<tr>
<td>Course Level</td>
<td>0.017</td>
<td>-1.131</td>
<td>1.638</td>
<td>-0.204</td>
<td>1.388</td>
<td>-0.523</td>
<td>2.814</td>
<td>0.216</td>
</tr>
<tr>
<td>Enjoyment of Grammar</td>
<td>-0.750</td>
<td>-0.796</td>
<td>-1.696</td>
<td>-0.556</td>
<td>-1.037</td>
<td>-0.927</td>
<td>-2.404</td>
<td>-0.690</td>
</tr>
</tbody>
</table>
To check the assumption of linearity, scatterplots of the relationships between ID measures and post-test scores were examined, both overall and by treatment condition, for each of the ID variables. According to Larson-Hall (2010), the linearity assumption “must be tested by examining the data visually” (p. 159). Both a regression line and a Loess line were superimposed on each graph. Since a Loess line considers small intervals of the data at a time, creating “a locally weighted running-line smoother” (p. 153), it can help to identify non-linear trends. The more closely the regression and Loess lines match for each scatterplot, the more confidence one can have that the linearity assumption has been met.

Homoscedasticity involves the residual variance not accounted for by a regression line. A regression balances error above and below a line so that the sum of the residuals (positive and negative) equals zero. The assumption of homoscedasticity is that these residual errors should not show a pattern; the variance of Y should be equal at all values of X. If the residuals do seem to show a non-random pattern (e.g., if greater variability in the residuals is found as the value of X increases), that serves as another indication that the relationship between the variables may not be linear. To check for violations of this assumption (both overall and for each group separately), studentized residuals were saved when running the regressions between the ID variables and post-test scores; then, a residuals plot was generated for each regression with the relevant ID measure on the X axis and the residuals on the Y axis. Two types of visual display were examined: Each residuals plot was checked for random scatter with constant spread across the graph, and the normal probability plot (also generated during the process) was checked to verify that the residuals clustered closely around the diagonal line.
Each of the ID measures’ characteristics will be discussed in turn with regard to the normality, linearity, and homoscedasticity assumptions. According to Howell (2002), when the assumptions of parametric (Pearson) correlation are violated, Spearman’s rho (ρ) can be used as a nonparametric alternative. The latter is also preferable for ordinal data, which is relevant for the ID measures which were rated on a Likert-style scale.

For the VPT (visual STM) and MLAT IV (grammatical sensitivity), the skewness and kurtosis ratios are acceptable, and the data look fairly normal in distribution (Appendices O1-O2). The groups’ variances also appear to be sufficiently similar, as confirmed by Levene’s tests. Deviations in the Loess lines compared to the straight regression lines on scatterplots showing the relationships between these IDs and learners’ post-test scores in each group do not seem to be a major cause for concern. As far as homoscedasticity is concerned, the residuals seem to pattern randomly, with relatively constant spread, except that there might be more data points and more unaccounted-for variance higher in the range of MLAT IV scores. Thus, parametric statistics seem acceptable for both of these ID measures, though nonparametric statistics seem worth considering for the MLAT IV.

On the MLAT V (rote memory), there appears to have been a ceiling effect: A substantial proportion of the learners in each group achieved the maximum score of 24. The data from all groups look negatively skewed (Appendix O2), and the skewness ratios for the No-Feedback and Trees-Feedback groups fall outside the range of -2 to 2. Furthermore, a statistically significant result for Levene’s test indicates that the variances cannot be considered equal; the No-Feedback group seems to show a greater amount of variance, while 5 learners in the Trees group are identified as outliers. The Right/Wrong-Feedback group has no outliers and a more
acceptable degree of skew, but still a relatively high skewness ratio. Regarding the linearity and homoscedasticity assumptions, the concern is not so much deviations in the Loess lines as the fact that so many more data points (and more variance in post-test scores) are located at the upper end of the MLAT V range; the residuals plots do not show constant spread. Considering that only 2 learners in the No-Feedback group were identified as low-scoring outliers, but several learners in that group actually scored below some of the learners in the Trees group who were identified as low-scoring outliers, it is difficult to justify excluding all (or only some) of them. Given these assumption violations, parametric statistics do not seem advisable for correlations with MLAT V scores.

For the test of sensitivity to ambiguity and knowledge of metalinguistic terminology (referred to in brief here as Ambiguity and Grammar) overall, the data for each group seem to be sufficiently normal in distribution for parametric statistics to be used (Appendix O3). Despite the fact that the median lines do not appear in the middle of the IQR boxes, the histograms do not look particularly skewed, and all of the skewness and kurtosis ratios are within range. Variances across groups are also fairly similar, as confirmed by Levene’s test. The scatterplots and residuals plots appear to meet the assumptions of linearity and homoscedasticity; spikes in subgroups’ Loess lines do not seem to be due to true curvilinearity, but rather possibly to the influence of small clusters of data points in otherwise sparse areas.

For the subtests of the Ambiguity and Grammar measure, however, more caution may be warranted. The formal tests for homogeneity, skewness, and kurtosis do not indicate any problems, the variances do indeed appear to be similar across groups, and each group has data points throughout the range of scores; however, the histograms do not seem to follow normal
curves very closely. On the Ambiguity subtest (Appendix O4), the skewness ratio for the No-
Feedback group is within the acceptable range, but relatively low at -1.856, and the Trees group
does not have as many scores as the others in the lower part of the range. Another important
issue to keep in mind is that some of the outliers on the Ambiguity subtest in particular may be
so in relation to the fact that their L1 was not English (i.e., construct-irrelevant variance). For
the Grammar subtest (Appendix O5), the No-Feedback and Trees-Feedback groups’ medians
appear toward the bottoms of their IQR boxes, and their distributions might be described as
somewhat bimodal. The scatterplots do not seem to have identifiable nonlinear patterns,
however, and the residuals plots seem sufficiently random. In sum, for the various subtests of
the Ambiguity and Grammar measure, parametric statistics may not be completely out of the
question, but nonparametric correlations may be more appropriate.

Considering the participants’ years of Japanese study and current Japanese course levels,
the conclusion is similar. For the groups separately, none of the skewness or kurtosis ratios are
outside the acceptable range for either variable; however, the data are positively skewed for
both variables when all participants are considered together, and the No-Feedback group’s data
might be bimodal for Years of Study (Appendix O6). Moreover, there may be differences in the
amount of variance across groups, with the Right/Wrong group showing less than the others for
Years of Study (no one in that condition had studied Japanese for more than 8 years), and the
No-Feedback group showing less than the others for Course Level (Appendix O7). In both cases,
the Levene value is relatively large, and it is statistically significant in the former case. Regarding
linearity and homoscedasticity for Years of Study, a few outliers who had studied Japanese for
longer than most of the other participants may exert disproportionate effects on the regression
lines. They also affect the spread of the residuals plots in that there are many more scores at the lower ends of the graphs. For Course Level, although there are no outliers, and although the scatterplots do not seem to show recognizable nonlinear patterns, there are fewer data points at the higher ends of the residuals plots, which therefore do not have constant spread. In any event, the ordinal nature of the Course Level variable means that nonparametric correlations may be more appropriate.

Finally, parametric statistics are not appropriate for the variables Number of Linguistics Courses (Appendix O8) or Enjoyment of Grammar (Appendix O9). Since most participants had taken only one or fewer linguistics courses, that variable shows floor effects, and all groups’ data are positively skewed, with several outliers. In fact, anyone who had taken 3 or more linguistics courses is considered an outlier. The Right/Wrong and Trees groups’ kurtosis ratios are outside the acceptable range, and, with regard to homoscedasticity, the residuals plots show many more data points at their lower ends. Though Enjoyment of Grammar does not suffer as much from these problems, it is ordinal in nature, making nonparametric correlations more appropriate.

Since parametric correlations were deemed truly appropriate for only one of the ID measures (the VPT), nonparametric (Spearman rank-order) correlations were used for all analyses involving IDs and post-test scores. Evaluations of the linearity and homoscedasticity assumptions for the relationships between IDs and delayed post-test scores are not reported here, but the decision to use nonparametric correlations holds for them as well.
Equivalence of experimental groups on individual-difference measures

As mentioned above, the ultimate goal for this set of research questions was to investigate the issue of aptitude-treatment interactions—that is, to find out whether certain abilities were more or less relevant than others under different instructional conditions (and, if so, which and why). This was pursued by examining the correlations between ID measures and test scores within each experimental group separately to see if the relationships differed in strength according to the type of treatment. Given the relatively low sample size in each group (between 25-29 participants) and, correspondingly, the chance that learners with particular strengths or weaknesses may not have been distributed randomly across groups, it seemed useful to know whether, at the outset of the study, there were differences between the groups in the learners’ ability profiles. For instance, did learners in the Right/Wrong condition happen to have greater grammatical sensitivity? Did learners in the Trees condition happen to have lower visual STM? As Larson-Hall (2010) points out, “Possibly the best that can be said... is that a lack of statistical difference between groups before experimental testing is ‘comforting’ (Howell, 2002, p. 500)” (p. 275); it is common practice for researchers to take such precautions, but also crucial to keep in mind that just because group differences are not found statistically does not mean that they do not exist. Nonetheless, since evidence of group differences would be important to keep in mind when interpreting the results, the practice was followed here.

A series of omnibus one-way ANOVAs, run separately for each ID measure, in fact found no statistically significant differences among the treatment groups for any of the IDs: visual STM (as measured by the Visual Patterns Test: $F(2, 76) = 3.015, p = 0.06$), grammatical sensitivity (MLAT Part IV: $F(2, 77) = 0.704, p = 0.50$), rote memory for linguistic material (MLAT
Part V: $F(2, 77) = 1.353, p = 0.26$), sensitivity to ambiguity in English (Ambiguity: $F(2, 77) = 2.118, p = 0.13$), sensitivity to ambiguity in English sentences containing reflexives (Reflexives: $F(2, 77) = 0.830, p = 0.44$), knowledge of metalinguistic terminology (Grammar: $F(2, 77) = 0.129, p = 0.88$), years of Japanese study ($F(2, 77) = 2.214, p = 0.13$), Japanese course level ($F(2, 77) = 1.144, p = 0.32$), number of linguistics courses ($F(2, 77) = 0.015, p = 0.99$), or enjoyment of grammar ($F(2, 75) = 2.814, p = 0.07$). Statistically significant differences likewise were not found when total scores on these measures, where relevant, were considered (MLAT Short-Form Total: $F(2, 77) = 1.944, p = 0.15$); Ambiguity and Grammar Total: $F(2, 77) = 0.380, p = 0.69$).

There were cases in which the $p$-values for these tests approached statistical significance: namely, for the learners’ visual STM and reported enjoyment of grammar.

Traditionally, post-hoc tests are employed to determine more specifically where any differences may lie only if an omnibus test has produced a statistically significant result; thus, their use would not be advised here. However, since correlation analyses (presented below) suggested that these IDs were more relevant to performance in the Trees condition than in the others, it may be worth noting that it was not in this condition where learners were rated higher.

Regarding visual short-term memory, LSD post-hoc tests located a statistically significant difference between the Right/Wrong and No-Feedback groups, with the Right/Wrong group performing better (marked in Table 5.6 with superscript ‘a’). As for the participants’ enjoyment of grammar, LSD post-hoc tests located a statistically significant difference between the Right/Wrong and Trees groups, with the Right/Wrong group reporting greater enjoyment (marked in Table 5.6 with superscript ‘b’).
Relationships between individual differences and test performance overall

In exploring overall relationships between individual differences and test performance regardless of treatment condition, it is crucial to keep in mind that a strong relationship between an ID and test scores among learners who engaged in one type of treatment could be washed out by a relationship in the opposite direction between the same ID and test scores among learners who engaged in a different type of treatment. That being the case, the results presented in this section are intended as a first look at the data, but by no means should they be interpreted as a definitive description of which individual differences were most strongly associated with a particular type of learning. A richer source of descriptions to support (indirect) inferences about learning processes will be found in the aptitude-treatment interactions.

Since this is a study of L2 learning, the most interesting relationships would be between IDs and pre-post gains and/or pre-delay gains. If not found for gains, however, (particularly considering that learners in the No-Feedback and Right/Wrong-Feedback conditions did not tend to show much improvement) it is also interesting to examine whether various IDs were related to performance on the pre-tests, whether they were then related to performance on the post- and delayed post-tests, and whether they were related to performance on the post-tests when controlling for pre-test scores (i.e., partialling out pre-test scores using partial correlations).

Because of the violations found in the assumptions for parametric statistics, as discussed above, nonparametric bivariate Spearman rank-order correlations were used to assess the relationships between IDs on the one hand and the learners’ gains and test scores on the other. PASW Statistics 18 does not provide a nonparametric version of partial correlations;
thus, the correlations between IDs and post-test scores with pre-test scores partialled out should be interpreted with the caveat that the ability to find statistically significant results may have been lessened due to assumption violations. Especially for those variables where assumptions were most clearly violated (e.g., number of linguistics courses, which was highly skewed, and enjoyment of grammar, which was ordinal), it is important not to over-interpret a lack of statistical significance.

It is also important to say a word about the alpha level. A common way of attempting to avoid Type I errors when multiple tests are run is to divide the alpha level by the number of tests. This is called a Bonferroni adjustment. In the current context, with exploratory correlations being run for each ID measure against each type of test (pre, post, and delay), adjusting in this conservative way would reduce the alpha level to such a strict threshold that very few results, if any, would reach statistical significance. The threat in that case would be rather of committing Type II errors (i.e., being overly cautious and assuming that no relationships exist even when they do). Pointing out that “the probability of Type II errors... [is] probably at least 20% or higher in most studies,” Larson-Hall (2010) argues that “[i]t doesn’t make a lot of sense to keep Type I errors at 5% while Type II errors are 20% or more!” (p. 101). Therefore, in the spirit of several other researchers (e.g., Kline, 2004; Murphy & Myors, 2004), she argues that “the alpha level should be set to $\alpha = .10$ in the social sciences” (p. 102). This more lenient practice will not be followed here, but the argument might at least lend support to a decision not to reduce the alpha level below 0.05. In any case, in interpreting correlations, effect sizes (the percentages of variance explained) are more revealing than $p$-values. Still, all results should be interpreted cautiously in light of the possibility that some are by chance.
Looking first at the correlations between ID measures and test performance for all of the participants considered together, Table 5.8a displays the Spearman rho (\(\rho\)) values, with \(p\)-values in parentheses, for the bivariate correlations between IDs and pre-test scores, IDs and post-test scores, and IDs and pre-post gains. The last column in the table shows the results of partial correlations between IDs and post-test scores with pre-test scores partialled out. Table 5.8b does the same for delayed post-test scores and gains. Following Cohen (1992), \(R^2\) values of .01, .09, and .25 are considered to be small, medium, and large effect sizes, respectively, indicating that 1%, 9%, and 25% of the variance is accounted for. Thus, the \(p\) values in the tables below can be considered small, medium, and large if they are above .10, .30, and .50.

The most obvious finding based on the results in Tables 5.8a and 5.8b is that while almost none of the ID measures show statistically significant correlations with learners’ gains from pre-test to post-test, gains from pre-test to delayed post-test, delayed post-test scores, or delayed post-test scores controlling for pre-test scores, nearly all of them show statistically significant small to moderate positive correlations with pre-test scores and post-test scores. Pre-test performance is predicted to a certain extent by the learners’ grammatical sensitivity (MLAT Part IV, \(\rho=.32\)) and MLAT total scores (\(\rho=.27\)), their sensitivity to ambiguity in English (\(\rho=.23\))—particularly on sentences involving reflexive pronouns (\(\rho=.29\))—and their Japanese course levels (\(\rho=.24\)), numbers of linguistics courses (\(\rho=.25\)), and enjoyment of grammar (\(\rho=.28\)). In some cases, the relationships with post-test scores are weaker (sensitivity to ambiguity with reflexives in English: \(\rho=.24\); number of linguistics courses: \(\rho=.18[ns]\); and enjoyment of grammar: \(\rho=.13[ns]\)); however, in many cases, the correlations appear to be slightly stronger in predicting post-test performance (MLAT IV: \(\rho=.35\); MLAT total: \(\rho=.28\); sensitivity to ambiguity: \(\rho=.24\); and,
especially, Japanese course levels: $\rho=.41$), and several IDs whose relationships with pre-test scores are not strong or reliable enough to reach statistical significance do predict post-test scores. This is the case for the learners’ visual short-term memory (VPT, $\rho=.30$), rote memory for linguistic material (MLAT V, $\rho=.30$), knowledge of metalinguistic terminology (Grammar, $\rho=.22$), and total scores on the test of sensitivity to ambiguity and knowledge of metalinguistic terminology ($\rho=.28$). The only ID variable not showing a statistically significant relationship with learners’ pre- or post-test scores is their years of Japanese study. Considered separately, the amount of variance in test scores explained by the IDs in the statistically significant relationships reviewed here ranges from 5% to 17%.

When pre-test scores are controlled for, there are statistically significant small to moderate positive correlations with post-test scores for learners’ visual short-term memory ($r=.30$), grammatical sensitivity (MLAT IV, $r=.29$), rote memory for language (MLAT V, $r=.24$), MLAT totals ($r=.27$), sensitivity to ambiguity ($r=.23$), total scores on the Ambiguity and Grammar measure ($r=.27$), and Japanese course levels ($r=.35$). The relationships with the Reflexives and Grammar subtests are not quite statistically significant, nor are the relationships with the learners’ years of Japanese study, number of linguistics courses, or enjoyment of grammar. Turning to the delayed post-test, only the learners’ course levels show a statistically significant relationship with delayed post-test scores ($\rho=.24$), and only their experience with linguistics courses shows a statistically significant relationship with delayed post-test scores when pre-test scores are controlled for ($r=.24$). These relationships are relatively small in magnitude, with only about 5% of the variance accounted for, but they are positive in direction. Tables 5.9-5.11 present correlation analyses for each experimental group separately, discussed below.
Table 5.8a. Bivariate Spearman correlations between ID measures and test scores, and partial correlations between IDs and post-test scores controlling for pre-test scores: All participants

<table>
<thead>
<tr>
<th>ID (pre N, post N, partial corr. df)</th>
<th>Pre-Test</th>
<th>Post-Test</th>
<th>Pre-Post Gain</th>
<th>Post Controlling for Pre</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPT (79, 78, 75)</td>
<td>.112 (.324)</td>
<td>.295** (.009)</td>
<td>.170 (.137)</td>
<td>.295** (.009)</td>
</tr>
<tr>
<td>MLAT IV (80, 79, 76)</td>
<td>.319** (.004)</td>
<td>.354** (.001)</td>
<td>.121 (.288)</td>
<td>.289** (.010)</td>
</tr>
<tr>
<td>MLAT V (80, 79, 76)</td>
<td>.216° (.054)</td>
<td>.300** (.007)</td>
<td>.122 (.286)</td>
<td>.236* (.037)</td>
</tr>
<tr>
<td>MLAT Total (80, 79, 76)</td>
<td>.267* (.016)</td>
<td>.281* (.012)</td>
<td>.083 (.465)</td>
<td>.267* (.018)</td>
</tr>
<tr>
<td>Reflexives Subtotal (80, 79, 76)</td>
<td>.290** (.009)</td>
<td>.238* (.035)</td>
<td>.028 (.810)</td>
<td>.169 (.139)</td>
</tr>
<tr>
<td>Ambiguity Subtotal (80, 79, 76)</td>
<td>.229* (.041)</td>
<td>.240* (.033)</td>
<td>.063 (.584)</td>
<td>.232* (.041)</td>
</tr>
<tr>
<td>Grammar Subtotal (80, 79, 76)</td>
<td>.112 (.323)</td>
<td>.223* (.048)</td>
<td>.078 (.494)</td>
<td>.193° (.090)</td>
</tr>
<tr>
<td>Ambiguity &amp; Grammar Total (80, 79, 76)</td>
<td>.182 (.106)</td>
<td>.279* (.013)</td>
<td>.101 (.376)</td>
<td>.265* (.019)</td>
</tr>
<tr>
<td>Years of Japanese Study (80, 79, 74)</td>
<td>.152 (.179)</td>
<td>.117 (.305)</td>
<td>.029 (.797)</td>
<td>.033 (.779)</td>
</tr>
<tr>
<td>Course Level (80, 79, 74)</td>
<td>.243* (.030)</td>
<td>.409** (.000)</td>
<td>.225* (.046)</td>
<td>.348**(.002)</td>
</tr>
<tr>
<td># Linguistics Courses (80, 79, 74)</td>
<td>.245* (.029)</td>
<td>.182 (.109)</td>
<td>.037 (.744)</td>
<td>.146 (.208)</td>
</tr>
<tr>
<td>Enjoyment of Grammar (78, 77, 74)</td>
<td>.282* (.012)</td>
<td>.130 (.261)</td>
<td>-.110 (.341)</td>
<td>.066 (.572)</td>
</tr>
</tbody>
</table>

Key: Spearman $p$-value ($p$-value)

* $p < 0.05$, ° $p < 0.10$
Table 5.8b. Bivariate Spearman correlations between ID measures and delayed test scores, and partial correlations between IDs and delayed test scores controlling for pre-test scores: All participants

<table>
<thead>
<tr>
<th>ID Measure</th>
<th>Delayed Post-Test</th>
<th>Pre-Delay Gain</th>
<th>Delay Controlling for Pre (df=73)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPT (78)</td>
<td>.016 (.889)</td>
<td>-.029 (.802)</td>
<td>.035 (.766)</td>
</tr>
<tr>
<td>MLAT IV (79)</td>
<td>.197° (.082)</td>
<td>-.006 (.961)</td>
<td>.164 (.160)</td>
</tr>
<tr>
<td>MLAT V (79)</td>
<td>.172 (.129)</td>
<td>-.023 (.841)</td>
<td>.125 (.284)</td>
</tr>
<tr>
<td>MLAT Total (79)</td>
<td>.124 (.278)</td>
<td>-.043 (.705)</td>
<td>.069 (.556)</td>
</tr>
<tr>
<td>Reflexives Subtotal (79)</td>
<td>.126 (.270)</td>
<td>-.128 (.263)</td>
<td>.068 (.562)</td>
</tr>
<tr>
<td>Ambiguity Subtotal (79)</td>
<td>.093 (.414)</td>
<td>-.095 (.407)</td>
<td>.062 (.600)</td>
</tr>
<tr>
<td>Grammar Subtotal (79)</td>
<td>.043 (.709)</td>
<td>-.028 (.810)</td>
<td>.029 (.807)</td>
</tr>
<tr>
<td>Ambiguity &amp; Grammar Total (79)</td>
<td>.087 (.444)</td>
<td>-.055 (.628)</td>
<td>.055 (.638)</td>
</tr>
<tr>
<td>Years of Japanese Study (79)</td>
<td>.021 (.856)</td>
<td>-.093 (.416)</td>
<td>-.027 (.819)</td>
</tr>
<tr>
<td>Course Level (79)</td>
<td>.237* (.036)</td>
<td>.023 (.842)</td>
<td>.171 (.142)</td>
</tr>
<tr>
<td># Linguistics Courses (79)</td>
<td>.168 (.139)</td>
<td>-.033 (.770)</td>
<td>.235* (.043)</td>
</tr>
<tr>
<td>Enjoyment of Grammar (77)</td>
<td>.058 (.615)</td>
<td>-.178 (.122)</td>
<td>.052 (.660)</td>
</tr>
</tbody>
</table>

Key: Spearman p-value (p-value)

* p < 0.05, ° p < 0.10
Table 5.9a. Bivariate Spearman correlations between ID measures and test scores, and partial correlations between IDs and post-test scores controlling for pre-test scores: No-Feedback condition

<table>
<thead>
<tr>
<th>ID (n, partial corr. df)</th>
<th>Pre-Test</th>
<th>Post-Test</th>
<th>Pre-Post Gain</th>
<th>Post Controlling for Pre</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPT (25, 22)</td>
<td>.043 (.839)</td>
<td>.297 (.150)</td>
<td>.300 (.146)</td>
<td>.279 (.187)</td>
</tr>
<tr>
<td>MLAT IV (25, 22)</td>
<td>.248 (.233)</td>
<td>.481* (.015)</td>
<td>.302 (.142)</td>
<td>.362o (.082)</td>
</tr>
<tr>
<td>MLAT V (25, 22)</td>
<td>-.055 (.796)</td>
<td>.162 (.440)</td>
<td>.172 (.412)</td>
<td>.166 (.438)</td>
</tr>
<tr>
<td>MLAT Total (25, 22)</td>
<td>.053 (.800)</td>
<td>.115 (.584)</td>
<td>.023 (.912)</td>
<td>.142 (.508)</td>
</tr>
<tr>
<td>Reflexives Subtotal (25, 22)</td>
<td>.090 (.668)</td>
<td>.269 (.193)</td>
<td>.182 (.383)</td>
<td>.254 (.231)</td>
</tr>
<tr>
<td>Ambiguity Subtotal (25, 22)</td>
<td>-.157 (.453)</td>
<td>.073 (.730)</td>
<td>.153 (.467)</td>
<td>.215 (.313)</td>
</tr>
<tr>
<td>Grammar Subtotal (25, 22)</td>
<td>.205 (.325)</td>
<td>.470* (.018)</td>
<td>.387o (.056)</td>
<td>.389o (.061)</td>
</tr>
<tr>
<td>Ambiguity &amp; Grammar Total (25, 22)</td>
<td>.074 (.725)</td>
<td>.414* (.040)</td>
<td>.374o (.065)</td>
<td>.398o (.054)</td>
</tr>
<tr>
<td>Years of Japanese Study (25, 21)</td>
<td>-.109 (.605)</td>
<td>-.261 (.207)</td>
<td>-.108 (.609)</td>
<td>-.240 (.271)</td>
</tr>
<tr>
<td>Course Level (25, 21)</td>
<td>.201 (.336)</td>
<td>.358o (.079)</td>
<td>.250 (.228)</td>
<td>.234 (.284)</td>
</tr>
<tr>
<td># Linguistics Courses (25, 21)</td>
<td>.142 (.499)</td>
<td>.491* (.013)</td>
<td>.370o (.068)</td>
<td>.327 (.128)</td>
</tr>
<tr>
<td>Enjoyment of Grammar (24, 21)</td>
<td>.259 (.222)</td>
<td>.042 (.845)</td>
<td>-.117 (.588)</td>
<td>-.069 (.755)</td>
</tr>
</tbody>
</table>

Key: Spearman $\rho$-value ($p$-value)

*p < 0.05, o $p < 0.10
Table 5.9b. Bivariate Spearman correlations between ID measures and delayed test scores, and partial correlations between IDs and delayed test scores controlling for pre-test scores: No-Feedback condition

<table>
<thead>
<tr>
<th>ID Measure</th>
<th>Delayed Post-Test</th>
<th>Pre-Delay Gain</th>
<th>Delay Controlling for Pre (df=21)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPT (25)</td>
<td>-.179 (.393)</td>
<td>-.070 (.738)</td>
<td>-.096 (.663)</td>
</tr>
<tr>
<td>MLAT IV (25)</td>
<td>.200 (.338)</td>
<td>.037 (.861)</td>
<td>.136 (.537)</td>
</tr>
<tr>
<td>MLAT V (25)</td>
<td>-.157 (.453)</td>
<td>-.093 (.659)</td>
<td>-.148 (.499)</td>
</tr>
<tr>
<td>MLAT Total (25)</td>
<td>-.230 (.269)</td>
<td>-.252 (.225)</td>
<td>-.268 (.216)</td>
</tr>
<tr>
<td>Reflexives Subtotal (25)</td>
<td>-.205 (.325)</td>
<td>-.377* (.063)</td>
<td>-.310 (.150)</td>
</tr>
<tr>
<td>Ambiguity Subtotal (25)</td>
<td>-.432* (.031)</td>
<td>-.327 (.111)</td>
<td>-.489* (.018)</td>
</tr>
<tr>
<td>Grammar Subtotal (25)</td>
<td>.158 (.451)</td>
<td>-.037 (.861)</td>
<td>.039 (.859)</td>
</tr>
<tr>
<td>Ambiguity &amp; Grammar Total (25)</td>
<td>-.082 (.695)</td>
<td>-.196 (.349)</td>
<td>-.259 (.233)</td>
</tr>
<tr>
<td>Years of Japanese Study (25)</td>
<td>-.267 (.197)</td>
<td>-.176 (.399)</td>
<td>-.164 (.454)</td>
</tr>
<tr>
<td>Course Level (25)</td>
<td>.395° (.051)</td>
<td>.198 (.343)</td>
<td>.403° (.057)</td>
</tr>
<tr>
<td># Linguistics Courses (25)</td>
<td>.284 (.169)</td>
<td>.226 (.277)</td>
<td>.346 (.106)</td>
</tr>
<tr>
<td>Enjoyment of Grammar (24)</td>
<td>-.133 (.537)</td>
<td>-.343° (.101)</td>
<td>-.233 (.284)</td>
</tr>
</tbody>
</table>

Key: Spearman p-value (p-value)

*p < 0.05, °p < 0.10
Table 5.10a. Bivariate Spearman correlations between ID measures and test scores, and partial correlations between IDs and post-test scores controlling for pre-test scores: Right/Wrong condition

<table>
<thead>
<tr>
<th>ID (n, partial corr. df)</th>
<th>Pre-Test</th>
<th>Post-Test</th>
<th>Pre-Post Gain</th>
<th>Post Controlling for Pre</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPT (28, 25)</td>
<td>.161 (.412)</td>
<td>.214 (.274)</td>
<td>.047 (.810)</td>
<td>.218 (.274)</td>
</tr>
<tr>
<td>MLAT IV (29, 26)</td>
<td>.545** (.002)</td>
<td>.138 (.475)</td>
<td>-.197 (.305)</td>
<td>.132 (.504)</td>
</tr>
<tr>
<td>MLAT V (29, 26)</td>
<td>.354° (.059)</td>
<td>.301 (.113)</td>
<td>.081 (.675)</td>
<td>.358° (.061)</td>
</tr>
<tr>
<td>MLAT Total (29, 26)</td>
<td>.457* (.013)</td>
<td>.226 (.238)</td>
<td>-.085 (.662)</td>
<td>.346° (.071)</td>
</tr>
<tr>
<td>Reflexives Subtotal</td>
<td>.315° (.096)</td>
<td>-.023 (.904)</td>
<td>-.159 (.411)</td>
<td>-.013 (.948)</td>
</tr>
<tr>
<td>Ambiguity Subtotal</td>
<td>.471** (.010)</td>
<td>.044 (.823)</td>
<td>-.213 (.268)</td>
<td>.117 (.552)</td>
</tr>
<tr>
<td>Grammar Subtotal</td>
<td>.103 (.595)</td>
<td>.182 (.345)</td>
<td>.082 (.672)</td>
<td>.157 (.424)</td>
</tr>
<tr>
<td>Ambiguity &amp; Grammar Total</td>
<td>.363° (.053)</td>
<td>.173 (.369)</td>
<td>-.058 (.765)</td>
<td>.185 (.345)</td>
</tr>
<tr>
<td>Years of Japanese Study</td>
<td>-.208 (.278)</td>
<td>.552** (.002)</td>
<td>.574** (.001)</td>
<td>.528** (.005)</td>
</tr>
<tr>
<td>Course Level (29, 25)</td>
<td>.007 (.969)</td>
<td>.560** (.002)</td>
<td>.483** (.008)</td>
<td>.615** (.001)</td>
</tr>
<tr>
<td># Linguistics Courses</td>
<td>.344° (.068)</td>
<td>.009 (.962)</td>
<td>-.162 (.402)</td>
<td>.175 (.383)</td>
</tr>
<tr>
<td>Enjoyment of Grammar</td>
<td>.148 (.453)</td>
<td>.115 (.559)</td>
<td>.016 (.937)</td>
<td>.083 (.679)</td>
</tr>
</tbody>
</table>

Key: Spearman p-value (p-value)  
*<sup>p</sup> < 0.05, °<sup>p</sup> < 0.10
Table 5.10b. Bivariate Spearman correlations between ID measures and delayed test scores, and partial correlations between IDs and delayed test scores controlling for pre-test scores: Right/Wrong condition

<table>
<thead>
<tr>
<th>ID Measure</th>
<th>Delayed Post-Test</th>
<th>Pre-Delay Gain</th>
<th>Delay Controlling for Pre (df=24)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPT (28)</td>
<td>-.156 (.428)</td>
<td>-.129 (.511)</td>
<td>-.087 (.673)</td>
</tr>
<tr>
<td>MLAT IV (29)</td>
<td>.049 (.799)</td>
<td>-.173 (.370)</td>
<td>.012 (.955)</td>
</tr>
<tr>
<td>MLAT V (29)</td>
<td><strong>.472</strong> (.010)</td>
<td>.205 (.286)</td>
<td><strong>.557</strong> (.003)</td>
</tr>
<tr>
<td>MLAT Total (29)</td>
<td>.169 (.381)</td>
<td>.001 (.994)</td>
<td>.282 (.163)</td>
</tr>
<tr>
<td>Reflexives Subtotal</td>
<td>.201 (.295)</td>
<td>-.003 (.987)</td>
<td>.261 (.198)</td>
</tr>
<tr>
<td>Ambiguity Subtotal</td>
<td>.208 (.279)</td>
<td>-.074 (.703)</td>
<td>.316 (.115)</td>
</tr>
<tr>
<td>Grammar Subtotal</td>
<td>-.176 (.360)</td>
<td>-.149 (.440)</td>
<td>-.179 (.383)</td>
</tr>
<tr>
<td>Ambiguity &amp; Grammar Total (29)</td>
<td>.081 (.675)</td>
<td>-.058 (.765)</td>
<td>.072 (.728)</td>
</tr>
<tr>
<td>Years of Japanese Study (29)</td>
<td>.405* (.029)</td>
<td><strong>.583</strong> (.001)</td>
<td>.448* (.022)</td>
</tr>
<tr>
<td>Course Level (29)</td>
<td>.257 (.178)</td>
<td>.277 (.146)</td>
<td>.269 (.183)</td>
</tr>
<tr>
<td># Linguistics Courses (29)</td>
<td>.093 (.631)</td>
<td>-.191 (.321)</td>
<td>.302 (.133)</td>
</tr>
<tr>
<td>Enjoyment of Grammar (28)</td>
<td><strong>.315</strong> (.103)</td>
<td>.163 (.407)</td>
<td>.303 (.132)</td>
</tr>
</tbody>
</table>

Key: Spearman p-value (p-value)

* p < 0.05, o p < 0.10
Table 5.11a. Bivariate Spearman correlations between ID measures and test scores, and partial correlations between IDs and post-test scores controlling for pre-test scores: Trees condition

<table>
<thead>
<tr>
<th>ID Measure</th>
<th>Pre-Test (n=26)</th>
<th>Post-Test (n=25)</th>
<th>Pre-Post Gain (n=25)</th>
<th>Post Controlling for Pre (df=22)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPT</td>
<td>.067 (.746)</td>
<td>.447* (.025)</td>
<td>.255 (.218)</td>
<td>.497* (.014)</td>
</tr>
<tr>
<td>MLAT IV</td>
<td>.171 (.403)</td>
<td>.558** (.004)</td>
<td>.378° (.063)</td>
<td>.453* (.026)</td>
</tr>
<tr>
<td>MLAT V</td>
<td>.334° (.095)</td>
<td>.433* (.031)</td>
<td>.040 (.849)</td>
<td>.280 (.185)</td>
</tr>
<tr>
<td>MLAT Total</td>
<td>.281 (.165)</td>
<td>.403* (.046)</td>
<td>.136 (.517)</td>
<td>.290 (.169)</td>
</tr>
<tr>
<td>Reflexives Subtotal</td>
<td>.421* (.032)</td>
<td>.378° (.062)</td>
<td>.045 (.829)</td>
<td>.287 (.174)</td>
</tr>
<tr>
<td>Ambiguity Subtotal</td>
<td>.287 (.154)</td>
<td>.486° (.014)</td>
<td>.188 (.369)</td>
<td>.397° (.055)</td>
</tr>
<tr>
<td>Grammar Subtotal</td>
<td>.097 (.636)</td>
<td>.160 (.445)</td>
<td>-.027 (.896)</td>
<td>.131 (.543)</td>
</tr>
<tr>
<td>Ambiguity &amp; Grammar Total</td>
<td>.146 (.477)</td>
<td>.332 (.105)</td>
<td>.109 (.603)</td>
<td>.265 (.210)</td>
</tr>
<tr>
<td>Years of Japanese Study</td>
<td>.634** (.001)</td>
<td>.048 (.819)</td>
<td>-.443* (.027)</td>
<td>-.263 (.215)</td>
</tr>
<tr>
<td>Course Level</td>
<td>.526** (.006)</td>
<td>.261 (.207)</td>
<td>-.158 (.451)</td>
<td>.019 (.930)</td>
</tr>
<tr>
<td># Linguistics Courses</td>
<td>.234 (.250)</td>
<td>-.002 (.994)</td>
<td>-.066 (.752)</td>
<td>.026 (.902)</td>
</tr>
<tr>
<td>Enjoyment of Grammar</td>
<td>.481** (.013)</td>
<td>.405* (.045)</td>
<td>-.019 (.929)</td>
<td>.317 (.131)</td>
</tr>
</tbody>
</table>

Key: Spearman p-value (p-value)

* p < 0.05, ° p < 0.10
Table 5.11b. Bivariate Spearman correlations between ID measures and delayed test scores, and partial correlations between IDs and delayed test scores controlling for pre-test scores: Trees condition

<table>
<thead>
<tr>
<th>ID</th>
<th>Delayed Post-Test</th>
<th>Pre-Delay Gain</th>
<th>Delayed Post Controlling for Pre</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPT</td>
<td>.267 (.197)</td>
<td>.108 (.608)</td>
<td>.192 (.369)</td>
</tr>
<tr>
<td>MLAT IV</td>
<td>.336* (.101)</td>
<td>.105 (.618)</td>
<td>.317 (.131)</td>
</tr>
<tr>
<td>MLAT V</td>
<td>.081 (.701)</td>
<td>-.246 (.236)</td>
<td>.127 (.555)</td>
</tr>
<tr>
<td>MLAT Total</td>
<td>.207 (.320)</td>
<td>-.075 (.722)</td>
<td>.118 (.584)</td>
</tr>
<tr>
<td>Reflexives Subtotal</td>
<td>.112 (.593)</td>
<td>-.238 (.252)</td>
<td>.157 (.465)</td>
</tr>
<tr>
<td>Ambiguity Subtotal</td>
<td>.181 (.385)</td>
<td>-.103 (.626)</td>
<td>.185 (.387)</td>
</tr>
<tr>
<td>Grammar Subtotal</td>
<td>.210 (.314)</td>
<td>.039 (.853)</td>
<td>.308 (.143)</td>
</tr>
<tr>
<td>Ambiguity &amp; Grammar Total</td>
<td>.249 (.230)</td>
<td>.024 (.910)</td>
<td>.281 (.184)</td>
</tr>
<tr>
<td>Years of Japanese Study</td>
<td>-.145 (.489)</td>
<td>-.535** (.006)</td>
<td>-.292 (.166)</td>
</tr>
<tr>
<td>Course Level</td>
<td>-.030 (.887)</td>
<td>-.378* (.062)</td>
<td>-.173 (.419)</td>
</tr>
<tr>
<td># Linguistics Courses</td>
<td>-.017 (.937)</td>
<td>-.139 (.508)</td>
<td>.116 (.591)</td>
</tr>
<tr>
<td>Enjoyment of Grammar</td>
<td>.146 (.487)</td>
<td>-.301 (.144)</td>
<td>.199 (.352)</td>
</tr>
</tbody>
</table>

Key: Spearman $\rho$-value ($p$-value)

*p < 0.05, *$p$ < 0.10
Relationships between individual differences and test performance by group

Tables 5.9a and 5.9b present the results of correlation analyses for the learners in the No-Feedback condition alone. With only a subset of the participants entering into these calculations, fewer statistically significant results might be expected due to the smaller sample size (n=25). However, if certain abilities were particularly relevant to performing well in this type of instructional condition (engaging in practice interpreting Japanese sentences without receiving any feedback as to the accuracy of one’s interpretations), then relationships between those ID variables and test performance might emerge more clearly.

In the No-Feedback group, none of the ID measures predict learners’ performance on the pre-test. Furthermore, none of the IDs’ relationships with gains from pre- to post-test or pre- to delayed post-test are statistically significant (although some approach this threshold). With performance on the post-test itself, however, some ID variables do show statistically significant correlations: grammatical sensitivity (MLAT IV, \( \rho = .48 \)), knowledge of metalinguistic terminology (Grammar, \( \rho = .47 \)), and number of linguistics courses taken (\( \rho = .49 \)). All of these can be considered strong positive relationships, accounting for 22-24% of the variance. The learners’ total scores on the Ambiguity and Grammar measure also show a statistically significant relationship with post-test scores (\( r = .41, \text{ns} \)), likely driven by the Grammar subtest.

None of the partial correlations with post-test scores controlling for pre-test scores are statistically significant, but the ones approaching significance are largely the same as those which are significant for post-test scores directly: grammatical sensitivity (\( r = .36, \text{ns} \)), knowledge of metalinguistic terminology (\( r = .39, \text{ns} \)), and the Ambiguity and Grammar total (\( r = .40, \text{ns} \)). The relationship with the learners’ numbers of linguistics courses (\( r = .33, \text{ns} \)) does not have a \( p \)-value
under .10 when pre-test scores are partialled out, but again, it is important to keep in mind the extreme skew in that variable’s distribution, which would tend to work against the power of parametric statistics to find a relationship.

It is on the delayed post-test that the first statistically significant negative correlations are found. Learners who scored higher on the subtest of sensitivity to ambiguity in English are likely to have scored lower on the delayed post-test ($\rho= -.43$). The partial correlation between sensitivity to ambiguity and delayed post-test scores controlling for pre-test scores is also statistically significant and strongly negative ($r= -.49$), accounting for 24% of the variance. Another variable which closely approaches statistical significance in its relationship with delayed post-test scores is the learners’ Japanese course level ($\rho=.40, p=.05$). The partial correlation in this case is similar ($r= .40, p=.06$). Also worth mentioning are the statistically non-significant but moderately sized negative relationships between the learners’ sensitivity to ambiguity in English sentences with reflexives and their pre-delay gains ($\rho= -.38, p=.06$) and between their enjoyment of grammar and pre-delay gains ($\rho= -.34, p=.10$).

Tables 5.10a and 5.10b display the results of analyses correlating IDs with test performance for learners in the Right/Wrong-Feedback condition. The fact that several IDs show statistically significant positive relationships with pre-test scores complicates matters—not in the sense that this is difficult to explain on its own, but rather considering the desire to be able to draw straightforward group comparisons. It makes sense that learners with greater grammatical sensitivity (MLAT IV, $\rho=.55$) and sensitivity to ambiguity ($\rho=.47$), for example, might perform better on a pre-test hypothetically also involving sensitivity to grammar and ambiguity in their L2, whereas an ID such as visual short-term memory ($\rho=.16, \text{ns}$) might not
emerge as a statistically significant predictor. The reason the relationships between IDs and pre-test scores make matters more complicated, however, is because one would hope to find similar relationships (or a similar lack of relationships) across all of the experimental groups on the pre-test, followed by different patterns of relationships on the post-test, after the learners had experienced different treatment conditions. The results attested here make it important to keep in mind, when interpreting post-test relationships and comparing them across groups, that learners in the Right/Wrong-Feedback group were already showing relationships between IDs and Japanese test scores at the outset of the study, before they had completed any activities that differed from those completed by other participants.

By the time of the post-test, the only ID variables to produce statistically significant correlations with test performance for learners in the Right/Wrong-Feedback condition are their years of Japanese study (\(\rho=.55\)) and their Japanese course levels (\(\rho=.56\)). Both of these are strong positive correlations, accounting for approximately 30% of the variance. When pre-test scores are partialled out, these are also the only ID variables to show statistically significant relationships with post-test performance (\(r=.53\) for Years of Study and \(r=.62\) for Course Level—28% and 38% of the variance explained, respectively). A partial correlation between post-test scores and the learners’ rote memory for linguistic material approaches statistical significance (MLAT V, \(r=.36, p=.06\)), as does one with the MLAT total (\(r=.35, p=.07\)). With regard to pre-post gains, again, the only statistically significant bivariate correlations are with the participants’ years of study (\(\rho=.57\)) and course levels (\(\rho=.48\)).

\[38\] Other ID measures whose relationships with pre-test scores approach statistical significance include the Ambiguity and Grammar total (\(\rho=.36, p=.05\)), memory for linguistic material (MLAT V, \(\rho=.35, p=.06\)), number of linguistics courses (\(\rho=.34, p=.07\)), and sensitivity to ambiguity in sentences with reflexives (\(\rho=.32, p=.10\)).
Interestingly, on the delayed post-test for learners in the Right/Wrong-Feedback group, rote memory for linguistic material (MLAT V, $\rho=.47$) is a strong and statistically significant predictor of performance, accounting for 22% of the variance, and this relationship is even stronger with pre-test scores partialled out ($r=.56$), accounting for 31% of the variance. The participants’ years of Japanese study remain a statistically significant predictor of delayed post-test scores ($\rho=.41$) and of gains from pre-test to delayed post-test ($\rho=.58$), whereas their course levels do not.

Finally, Tables 5.11a and 5.11b present the results for learners in the Trees-Feedback condition. On the pre-test, the learners’ years of Japanese study ($\rho=.63$), Japanese course levels ($\rho=.53$), enjoyment of grammar ($\rho=.48$), and sensitivity to ambiguity in English sentences with reflexives ($\rho=.42$) show statistically significant correlations with performance. These are fairly strong correlations; the ID measures, considered independently, account for 18% to 40% of the variance in pre-test scores. By the time of the post-test, however, apart from enjoyment of grammar, these IDs no longer show statistically significant relationships with test performance. Instead, the individual differences most strongly predictive of post-test scores are the learners’ visual short-term memory (VPT, $\rho=.45$), grammatical sensitivity (MLAT IV, $\rho=.56$), rote memory for linguistic material (MLAT V, $\rho=.43$), MLAT totals ($\rho=.40$), general sensitivity to ambiguity in English ($\rho=.49$), and enjoyment of grammar ($\rho=.41$). The post-test relationship with learners’ sensitivity to ambiguity in English sentences with reflexives approaches statistical significance ($\rho=.38$, $p=.06$). On the delayed test, only the relationship between grammatical sensitivity and test performance even comes close to approaching statistical significance ($\rho=.34$, $p=.10$).
As far as pre-post and pre-delay gains are concerned for the Trees group, only two relationships with IDs are statistically significant, and they are fairly strong negative ones involving the learners’ years of Japanese study. The longer participants had been studying Japanese, the less improvement they showed from pre- to post-test ($\rho= -.44$) and from pre- to delayed post-test ($\rho= -.54$). In these cases, the strong positive relationship between length of study and pre-test scores should be kept in mind. A moderate positive relationship between grammatical sensitivity and pre-post gains ($p=.38$, $p=.06$) and a moderate negative relationship between learners’ course levels and pre-delay gains ($p=.38$, $p=.06$) also approach statistical significance. Looking at the relationships between IDs and post-test scores with pre-test scores controlled for, the partial correlations show two statistically significant relationships—with visual short-term memory ($r=.50$) and grammatical sensitivity ($r=.45$)—and one which approaches statistical significance—with sensitivity to ambiguity in English ($r=.40$, $p=.06$).

In a way, the pattern of results for the Trees group seems complementary to the pattern in the Right/Wrong group. Whereas the Right/Wrong group started out showing relationships between pre-test scores and sensitivity to grammar/ambiguity scores, then later showed relationships between post-test scores and their years and levels of Japanese study, the Trees group started out showing relationships between pre-test scores and their years and levels of Japanese study, then later showed relationships between post-test scores and sensitivity to grammar/ambiguity. (Of course, this comparison is not exhaustive; the Trees group additionally showed relationships involving enjoyment of grammar and visual short-term memory.)
Figure 5.3. Scatterplots of relationships between VPT and pre-/post-test scores, with a separate regression line for each group.
Figure 5.4. Scatterplots of relationships between MLAT IV and pre-/post-test scores, with a separate regression line for each group.
Figure 5.5. Scatterplots of relationships between MLAT V and pre-/post-test scores, with reference lines at 80% on each measure.
As another perspective on these data which allows for a consideration of all of the groups at once, Figures 5.3-5.5 display scatterplots of the relationships between selected ID measures and post-test scores. On some of these scatterplots, a separate regression line is drawn to represent the direction and strength of the relationship for each group. Moreover, each data point is marked according to group membership, using small circles (O) for the No-Feedback group, exes (X) for the Right/Wrong group, and inverted Ys (resembling mini tree diagrams) for the Trees group. It is important to note that these regression lines and R² values do not correspond exactly to the results of the nonparametric Spearman correlations presented earlier, which are less sensitive to outliers. However, the lines do represent the general nature of the relationships, and the goal here is to gain a visual sense of how all of the data patterned.

In Figure 5.3, it is straightforward to compare the experimental groups with respect to the associations shown between visual short-term memory and scores on the pre-test versus on the post-test. On the pre-test, as might be expected, all three groups show similarly flat relationships between visual STM and Japanese test scores. In contrast, by the time of the post-test, there are clearer differences among the groups; visual STM appears to be most strongly correlated with the test performance of learners in the Trees group, who were given visual diagrams as feedback during the treatments.

In Figure 5.4, which displays the relationships between grammatical sensitivity (MLAT IV scores) and test performance, it is interesting to try to ascertain what might be producing the groups’ different patterns of correlations. Looking more closely at the No-Feedback and Trees groups, they show relatively weak, statistically non-significant positive correlations on the pre-test, which then become strong and statistically significant on the post-test; this is relatively
straightforward. In the Right/Wrong group, in contrast, there is a strong, statistically significant relationship between grammatical sensitivity and pre-test performance, which is then negligible on the post-test. Here, it is helpful to trace subsets of learners. To facilitate this, rectangles are positioned around learners in the Right/Wrong group (the data points marked X) who scored roughly in the middle versus roughly at the top of the MLAT IV range. Interestingly, the middle-MLAT rectangle moves up in Japanese test scores from pre- to post-test, whereas the top-MLAT rectangle moves down. In other words, it may be that learners with average grammatical sensitivity tended to improve on the post-test after receiving simple acceptance/rejection feedback, whereas learners with higher grammatical sensitivity tended to perform worse after receiving this sort of feedback; it does not seem to be a case of ceiling effects, whereby learners with high grammatical sensitivity simply had no room to improve on the post-test.

Regarding the MLAT V scatterplots in Figure 5.5, which show relationships between test scores and rote memory for linguistic material, perhaps the most salient feature is the concentration of data points toward the right of the graph, illustrating a ceiling effect on the memory measure, with a large proportion of learners achieving the maximum score. In this case, possibly more instructive than inspecting regression lines is superimposing reference lines and inspecting the data points in quadrants. Placing a reference line at approximately 80% on each measure, what one sees in looking at the upper left and right quadrants is that, with the exception of one learner in the No-Feedback group, it is only participants who scored above 80% on the MLAT V who also scored above 80% on the post-test. For the one learner in the No-Feedback group who scored low on MLAT V but relatively high on the post-test, memory ability might have been less relevant anyway since no feedback was provided in that condition.
Appendices P1-P8 contain similar scatterplots, illustrating the relationships between test scores and the rest of the ID variables (i.e., sensitivity to ambiguity, knowledge of metalinguistic terminology, years of Japanese study, Japanese course level, number of linguistics courses taken, and enjoyment of grammar). These graphs likewise help to shed light on how learners from each of the different experimental groups contributed to ID/test correlations of different directions and strengths on the pre-test versus the post-test.

**An additional perspective: Multiple regression**

Before moving on to the next chapter, which presents a more detailed interpretation of the results in relation to the research questions, a word should be said about another possible statistical approach to making sense of the relationships between individual differences and test performance across different treatment groups: multiple regression. Due to limitations in sample size and the violations of statistical assumptions already noted, the full potential of this technique is beyond reach, and it did not seem advisable to use it as the main approach to analysis, but it can provide another perspective on what has already been reported.

When several predictor variables (e.g., IDs) have been measured, as in the present study, it can be illuminating to investigate which of them are able to account for the most variance on an outcome measure independently of the others. In a standard multiple regression, the main research questions are (1) how much total variance in a response variable can be explained by a combination of predictor variables, and (2) how much each predictor variable contributes uniquely to the relationship. In this case, areas of overlap between the predictor variables are excluded (Larson-Hall, 2010, p. 179), and in providing a semipartial correlation coefficient for
each predictor variable, researchers can indicate their relative importance (p. 205). In a sequential multiple regression, on the other hand, researchers can more proactively emphasize the importance of certain predictor variables by controlling the order in which they are entered into the analysis. In that case, areas of overlap are not excluded; the main research question is whether each subsequent variable entered after a previous one can further increase the amount of variance explained (p. 180).

The assumptions behind multiple regression are largely the same as those for bivariate correlations, which have already been discussed. Apart from linearity and homogeneity of variances, multiple regression requires that the predictor variables not be highly intercorrelated, that the sample size be adequate, and that there be no outliers. Concerning multicollinearity, Tabachnick and Fidell (2001) state that correlations between the predictor variables should not be greater than $r=.70$. As shown in Table 5.12 (which presents Spearman rho coefficients for correlations between the ID variables discussed above), the only relationships stronger than .70 are those between subtotal and total scores of the same measure, which would not be entered into the same regression model anyway.
### Table 5.12. Intercorrelations among individual-difference variables

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MLAT IV</td>
<td>.207°</td>
<td>1 (80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT V</td>
<td>.139</td>
<td>.350**</td>
<td>1 (80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT Total</td>
<td>.164</td>
<td>.753**</td>
<td>.570**</td>
<td>1 (80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reflexives Subtotal</td>
<td>.009</td>
<td>.314**</td>
<td>.312**</td>
<td>.395**</td>
<td>1 (80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambiguity Subtotal</td>
<td>.067</td>
<td>.454**</td>
<td>.393**</td>
<td>.544**</td>
<td>.858**</td>
<td>1 (80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammar Subtotal</td>
<td>.032</td>
<td>.430**</td>
<td>.235*</td>
<td>.338**</td>
<td>.284*</td>
<td>.293**</td>
<td>1 (80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambiguity &amp; Grammar</td>
<td>.069</td>
<td>.550**</td>
<td>.382**</td>
<td>.554**</td>
<td>.676**</td>
<td>.756**</td>
<td>.825**</td>
<td>1 (80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Japanese Study</td>
<td>-.006</td>
<td>-.100</td>
<td>-.027</td>
<td>.095</td>
<td>.183</td>
<td>.071</td>
<td>-.056</td>
<td>.036</td>
<td>1 (80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Course Level</td>
<td>.229*</td>
<td>.049</td>
<td>.072</td>
<td>.059</td>
<td>.023</td>
<td>-.100</td>
<td>.110</td>
<td>.003</td>
<td>.356**</td>
<td>1 (80)</td>
<td>(80)</td>
<td>(80)</td>
<td>(78)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Linguistics Courses</td>
<td>-.025</td>
<td>.136</td>
<td>.141</td>
<td>-.017</td>
<td>.303**</td>
<td>.252*</td>
<td>.194°</td>
<td>.237*</td>
<td>-.068</td>
<td>.028</td>
<td>1 (80)</td>
<td>(78)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enjoyment of Grammar</td>
<td>-.004</td>
<td>.198°</td>
<td>.163</td>
<td>.003</td>
<td>.133</td>
<td>.157</td>
<td>.229*</td>
<td>.206°</td>
<td>.019</td>
<td>-.007</td>
<td>.175</td>
<td>1</td>
<td>(78)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Correlations are shown at the bottom left of the table. The top right shows the sample sizes in parentheses. Key: **p < 0.01, *p < 0.05, °p < 0.10
As alluded to above, sample size is a critical issue here. Many different rules of thumb have been proposed regarding the number of participants required for each predictor variable included in a multiple regression: 10 (Howell, 2002); 15 (Stevens, 2002); 30 (Porte, 2002); 50+8m, where ‘m’ equals the number of predictor variables (Tabachnick & Fidell, 2001); and a variety of other recommendations which incorporate estimated effect sizes and power (summarized in Larson-Hall, 2010, p. 184). Even if the most relaxed recommendation of 10 participants per predictor variable is followed here, in a study with 80 participants, not all of the ID measures can be included. More importantly, since the main interest is in which variables predict performance in the different experimental conditions considered separately, with fewer than 30 participants per group, only 2 ID measures might be expected to show statistically significant contributions in each of those analyses.

For the Right/Wrong-Feedback condition (n=29), in which only 2 ID variables had shown statistically significant bivariate correlations with post-test scores (Years of Study \( \rho=.55 \) and Course Level \( \rho=.56 \)), a multiple regression analysis including them as predictor variables produces findings similar to those reported above. The model is statistically significant, with an R value of .64, meaning that 41% of the variance in post-test scores within that group can be explained by variance in the participants’ years of Japanese study and course levels \( R^2=.41, \) Adjusted \( R^2=.36, \) Std. Error: 10.55). Table 5.13 presents the unstandardized regression coefficients (B), 95% confidence intervals, and squared semipartial correlations \( (sr^2) \) for the terms in the model, and indicates that each term is statistically significant. Each \( sr^2 \) value expresses “the unique contribution of each variable to the overall \( R^2 \)” (Larson-Hall, 2010, p. 198). If a third predictor variable, MLAT V (rote memory), is added to the model, the R value
increases to .69 (R²=.47, Adjusted R²=.40), but the term is not statistically significant (sr²=.25, B=1.288, p=.10), confirming the results of the previous correlation analyses.  

Table 5.13. Multiple regression model predicting post-test scores in the Right/Wrong group

<table>
<thead>
<tr>
<th></th>
<th>sr²</th>
<th>B</th>
<th>Std. Error</th>
<th>t</th>
<th>p</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>.37542</td>
<td>6.837</td>
<td>5.491</td>
<td>.000</td>
<td>23.461</td>
<td>51.622</td>
<td></td>
</tr>
<tr>
<td>Course Level</td>
<td>.419</td>
<td>4.184</td>
<td>1.539</td>
<td>2.719</td>
<td>.012</td>
<td>1.014</td>
<td>7.353</td>
</tr>
<tr>
<td>Years of Study</td>
<td>.344</td>
<td>3.460</td>
<td>1.554</td>
<td>2.227</td>
<td>.035</td>
<td>.260</td>
<td>6.660</td>
</tr>
</tbody>
</table>

In performing a multiple regression analysis for the Trees condition, a problem is posed by the fact that there are 5 statistically significant bivariate correlations between IDs and post-test scores (VPT, MLAT IV, MLAT V, Ambiguity, and Enjoyment of Grammar), with a sample size that allows for only 2 predictor variables to be entered into the model. Moreover, possibly due to intercorrelations between the predictor variables, MLAT IV (which correlates the most strongly with post-test scores) cannot be paired with either of the IDs which are the next most strongly correlated with post-test scores (VPT, Ambiguity) and still have both terms in the regression model be statistically significant; inclusion of MLAT IV produces an adequate model only when paired with Enjoyment of Grammar (R=.63, R²=.39). Meanwhile, a model including only VPT and Ambiguity, while excluding MLAT IV, is also statistically significant and accounts for (very slightly) greater variance (R=.66, R²=.44). Tables 5.14 and 5.15 display these models.

39 According to guidelines provided by Larson-Hall (2010, p. 196), the residuals statistics were all within acceptable ranges (i.e., standardized residuals between +/-3, Cook’s distance less than 1, Mahalanobis distance less than 11), and the VIF (variance inflation factor) multicollinearity values, which should be less than 5, were around 1.
Table 5.14. Multiple regression model predicting post-test scores in the Trees group

<table>
<thead>
<tr>
<th></th>
<th>$\text{sr}^2$</th>
<th>B</th>
<th>Std. Error</th>
<th>$t$</th>
<th>$p$</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>33.892</td>
<td>11.498</td>
<td>2.948</td>
<td>.007</td>
<td></td>
<td>10.047</td>
<td>57.737</td>
</tr>
<tr>
<td>MLAT IV</td>
<td>.433</td>
<td>1.019</td>
<td>.392</td>
<td>2.601</td>
<td>.016</td>
<td>.207</td>
<td>1.831</td>
</tr>
<tr>
<td>Enjoyment of Grammar</td>
<td>.403</td>
<td>3.978</td>
<td>1.643</td>
<td>2.422</td>
<td>.024</td>
<td>.571</td>
<td>7.385</td>
</tr>
</tbody>
</table>

Table 5.15. Another multiple regression model predicting post-test scores in the Trees group

<table>
<thead>
<tr>
<th></th>
<th>$\text{sr}^2$</th>
<th>B</th>
<th>Std. Error</th>
<th>$t$</th>
<th>$p$</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-2.636</td>
<td>18.325</td>
<td>-.144</td>
<td>.887</td>
<td>.887</td>
<td>-40.640</td>
<td>35.368</td>
</tr>
<tr>
<td>Ambiguity Subtotal</td>
<td>.476</td>
<td>2.664</td>
<td>.893</td>
<td>2.983</td>
<td>.007</td>
<td>.812</td>
<td>4.517</td>
</tr>
<tr>
<td>VPT</td>
<td>.468</td>
<td>6.596</td>
<td>2.248</td>
<td>2.934</td>
<td>.008</td>
<td>1.934</td>
<td>11.258</td>
</tr>
</tbody>
</table>

40 For the sake of completeness, the process of conducting these analyses for the Trees group will be described in more detail. Participant 59 did not complete the immediate post-test, and Participant 107 was an outlier on MLAT IV. Thus, they were excluded from initial analyses. Participants 1, 37, 46, 65, and 83 were also excluded from correlations with MLAT V scores since they were outliers on that measure. Looking then at the correlations between IDs and post-test scores, the following coefficients were found (presenting Pearson since multiple regression is parametric): MLAT IV ($r=.56^{**}$, $p=.005$), VPT ($r=.49^*$, $p=.02$), Ambiguity ($r=.47^*$, $p=.02$), Enjoyment of Grammar ($r=.47^*$, $p=.02$), and MLAT V ($r=.27$, $p=.26$). Since the correlation between MLAT V and post-test scores was not statistically significant when the outliers were excluded, it was not a candidate for inclusion in a multiple regression. The decision was made to include MLAT IV in the model first since its correlation with post-test scores was the highest. VPT was chosen next since it showed the next strongest correlation with post-test scores and was less intercorrelated with MLAT IV scores than the Ambiguity subtotal was. (The correlations between MLAT IV and the other ID variables were as follows: Ambiguity [$r=.51^{**}$, $p=.01$], VPT [$r=.43^*$, $p=.03$], Enjoyment of Grammar [$r=.26$, $p=.23$].) As it turned out, in the multiple regression model with MLAT IV and VPT ($R=.62$, $R^2=.38$), the MLAT IV term was statistically significant ($sr^2=.38$, $B=1.202$, $p=.04$), but the VPT term was not ($sr^2=.27$, $B=4.416$, $p=.13$). Thus, a model with MLAT IV and the Ambiguity subtotal was tested, with similar results ($R=.60$, $R^2=.36$; MLAT IV $sr^2=.37$, $B=1.215$, $p=.047$; Ambiguity $sr^2=.22$, $B=1.404$, $p=.23$). When MLAT IV and Enjoyment of Grammar were tried together, both terms were significant, as described in the text and shown in Table 5.14. However, considering that MLAT IV and the Ambiguity subtotal might tap similar abilities, it was decided to check a model with the Ambiguity subtotal and VPT as the 2 predictor variables, including Participant 107 again since he was not an outlier on either of those measures. The correlation between those IDs was negligible ($r=-.02$, $p=.94$). This model is described in the main text and in Table 5.15. When Enjoyment of Grammar was then added, so that 3 predictor variables were included, the $R$ value increased a bit ($R=.68$, $R^2=.47$), but the additional term was not statistically significant ($sr^2=.17$, $B=1.838$, $p=.31$), as might be expected with such a small $n$. 

300
Without a principled way of deciding between these models, the results are not particularly enlightening beyond what can already be gathered from the canonical correlations. Larson-Hall (2010) explains: “In general, a simpler model is better than a more complicated one, if they have the same explanatory value. A model with less error will be better than one with more error;” however, “whatever model you use to fit your data, there will be error” (p. 200). Or, as Crawley (2007) puts it, “All models are wrong” (p. 339). In the present case, with little difference between the models in terms of explanatory value, it seems the main ‘finding’ is that since certain ID variables are intercorrelated, they “appear to contribute very little to the overall regression equation in spite of being highly correlated with the response variable” (Larson-Hall, 2010, p. 179). The existence of these intercorrelations is interesting, but can also be discovered more directly, as exemplified in Table 5.12, which presents Spearman rank correlations between pairs of ID measures (without excluding outliers). Worth mentioning from that table are the following trends:

- The only ID variables which show statistically significant correlations with the learners’ Japanese course level are their years of Japanese study ($\rho=.36$) and visual STM ($\rho=.23$). For years of study and visual STM, these are the only statistically significant correlations with any of the other IDs.

- The number of linguistics courses taken, on the other hand, while not showing statistically significant correlations with MLAT scores, does show statistically significant positive correlations with participants’ sensitivity to ambiguity in English sentences ($\rho=.25$), including those with reflexives ($\rho=.30$). The relationship with their knowledge of metalinguistic terminology is weak, but approaches statistical significance ($\rho=.19$).
• The participants’ enjoyment of grammar shows a statistically significant positive correlation with their knowledge of metalinguistic terminology ($\rho=.23$), but with none of the other IDs (though it approaches statistical significance for the relationship with MLAT IV grammatical sensitivity scores, $\rho=.20$).

• Parts IV and V of the MLAT are moderately positively correlated with each other ($\rho=.35$), as are the Ambiguity and Grammar subtests with each other ($\rho=.29$).

• The Ambiguity and Grammar total and the MLAT are correlated with each other at $\rho=.55$, with the Ambiguity subtest showing a stronger correlation with the MLAT ($\rho=.54$) than the Grammar subtest does ($\rho=.34$). MLAT IV (grammatical sensitivity) shows relatively strong and statistically significant positive correlations with the sensitivity to ambiguity subtest ($\rho=.45$) and the knowledge of metalinguistic terminology subtest ($\rho=.43$), as well as with the total score on the Ambiguity and Grammar measure ($\rho=.55$). These relationships are stronger than those found between MLAT V (rote memory for language) and the Ambiguity ($\rho=.39$) and Grammar ($\rho=.24$) subtests and total ($\rho=.38$).

None of these relationships are especially surprising. The point in presenting them, rather, is that they may be of use in relation to theories of aptitude complexes (e.g., to understand which sorts of abilities and experiential variables might be related) and in practical terms (e.g., in developing measures to test such abilities efficiently). If learners’ scores on a grammar test and their reported enjoyment of grammar are very highly intercorrelated, for instance, and each of them is able to predict a similar amount of variance in another outcome measure, one might feel justified in taking a few seconds simply to ask a learner how much s/he enjoys grammar rather than taking substantially more time to administer the grammar test.
CHAPTER 6

DISCUSSION

Overview of research questions and results

The research reported on in this dissertation was motivated by several aims: to contribute to the literature on the effectiveness of more versus less explicit types of feedback in computer-assisted language learning; to attempt a preliminary test of one of S. E. Carroll’s (2001) proposed constraints on the effectiveness of feedback; to investigate which of a variety of individual differences might be related to L2 performance under different conditions of exposure; and to explore how aptitude-treatment interactions might enable finer-grained and more plausibly explanatory interpretations of the results of an experiment in which different types of feedback were compared. For the linguistic target of this study (interpretations of the Japanese reflexive *zibun*), relevant structural characteristics are invisible in the sense that they are not represented directly in surface morphosyntax, and it has been suggested that, for speakers of English, *zibun* presents a poverty-of-the-stimulus learning situation in which both positive and negative evidence may theoretically be required due to differences between English and Japanese. Three main research questions guided the study: (1) whether providing tree diagrams as feedback could help learners to interpret sentences with *zibun* more accurately (both over time and in comparison to other conditions of interpretation practice); (2) which individual differences might predict success (particularly with this novel type of feedback); and (3) whether individual differences would be differentially related to performance according to the type of interpretation practice the learners engaged in.
In actuality, the answers to these questions are inextricably intertwined. Individual differences can have an influence on whether treatment effects are found (i.e., a treatment may work quite well, and quite a bit better than other types of treatment, but only for certain types of learners), and differences in treatments can have an influence on which individual differences are relevant to success (i.e., an ‘aptitude’ associated with better performance under one condition of exposure might even be detrimental under another).

In the present study, it was found that only the group of learners who received tree diagrams as feedback clearly demonstrated sustained improvement in their interpretations of sentences containing reflexives. This is important information, but it is equally crucial to point out that group averages can reveal one pattern while concealing others: Not all of the participants in the Trees-Feedback condition achieved higher scores following the treatment, some of the participants in the other conditions did achieve higher scores, and performance on the tests was differentially related to individual learner variables according to treatment condition. The post-test scores of participants who had been provided with tree diagrams were related to their visual short-term memory and grammatical sensitivity, and possibly also to their rote memory for linguistic material, sensitivity to ambiguity, and reported enjoyment of grammar. Meanwhile, for participants who had simply been informed about the correctness of each response without being shown tree diagrams, post-test scores were related to the number of years they had been studying Japanese and to their current Japanese course levels. In a treatment condition which provided no feedback on sentence interpretations, post-test scores may have been related to the participants’ grammatical sensitivity, knowledge of metalinguistic terminology, and numbers of linguistics courses taken.
In other words, while it is certainly interesting to know whether one type of treatment might tend to work better than others overall, and while it is certainly interesting to examine whether particular individual differences predict L2 performance in general, it can be much more illuminating to explore how the characteristics of different treatment conditions can make various individual differences more or less relevant to success—that is, to investigate what works for whom (and when, and why). “Aptitude” is necessarily aptitude for something. The concept is empty without a situation in which abilities (or knowledge or predispositions) can be put to use in some way toward some purpose. Considering the wide variety of situations of L2 exposure which exist (and which we could create), there are possibly quite a lot of abilities that SLA researchers could investigate in order to reveal which emerge as aptitudes in different contexts. I will argue below that knowledge of aptitude-treatment interactions has great potential to advance the development of the field of SLA through improving how we design our studies, deepening what we are able to learn about the intricate processes of language acquisition, and ultimately contributing to the optimization of language instruction.

_Effectiveness of feedback (Research Question 1)_

In this experiment, adult English-speaking learners of L2 Japanese were randomly assigned to different treatment conditions and pre-tested on their interpretations of the reflexive _zibun_. Each participant then engaged in one of three types of learning activity: simple sentence interpretation practice without feedback (referred to as the No-Feedback condition), sentence interpretation practice with feedback indicating the correctness of each response (the Right/Wrong-Feedback condition), or sentence interpretation practice with feedback indicating
the correctness of each response and then providing a tree diagram to illustrate the underlying structure of any sentence the participant had interpreted incorrectly (the Trees-Feedback condition). Efforts were made to design the testing and treatment materials in such a way that rules of thumb the participants might develop based on, for example, case-markings or word order, would be contradicted by the feedback they received, if any, and the only way in which they could therefore perform with 100% accuracy would be if they were sensitive to the pertinent underlying structural relationships.

As reported in Chapter 5, no statistically significant differences among the groups were found on the pre-test. Performing a repeated-measures ANOVA for each treatment condition separately, it was found that only the Trees-Feedback group demonstrated improvement over the course of the tests (i.e., a statistically significant effect for Time across the pre-, post-, and delayed post-tests). Paired t-tests for that group also revealed statistically significant score increases, with large effect sizes, from the pre-test to the post-test and from the pre-test to the delayed post-test, suggesting that participants in the Trees condition not only began to interpret sentences with zibun more accurately, but tended to maintain their accuracy gains. (A paired t-test comparing their post-test and delayed post-test scores did not indicate a statistically significant decrease.) In the Right/Wrong-Feedback condition, in contrast, no clear evidence of improvement was found, and in the No-Feedback condition there was only limited evidence of improvement: A RM ANOVA comparing the No-Feedback group’s performance across all three tests did not produce a statistically significant result, but a paired t-test comparing their pre- and post-test scores was statistically significant, with a small effect size. By
the time of the delayed post-test, however, according to another paired t-test, the No-Feedback group no longer exhibited higher test scores compared to pre-test levels.

Evidence of a difference in learning gains among the treatment conditions would have been provided by a statistically significant Time-by-Treatment interaction effect in a single RM ANOVA comparing the groups’ pre-, post-, and delayed post-test performances directly, and it is important to emphasize that this was not found when all of the participants were included. That is, based solely on the set of results reviewed above, even though the Trees-Feedback group was the only one to show sustained improvement, it is not strictly possible to say that they outperformed the others with regard to the amount of improvement shown over time.

Interestingly, though, in a result that anticipates the issue of aptitude-treatment interactions, this conclusion differs from what might have been concluded if the experiment had been conducted in a more limited set of Japanese programs. When exactly the same statistics are run including test scores only from participants at Georgetown University (GU) and the University of Maryland (UMD), where data were collected first, even though the sample size is smaller, the Time-by-Treatment interaction is on the cusp of statistical significance \( (F(3.43, 92.47) = 2.555, p = 0.05, \eta_p^2 = 0.09) \), suggesting that the groups’ trajectories of pre- to post- to delayed post-test performance differed from each other. The effect for Time is also statistically significant \( (F(1.71, 92.47) = 13.27, p < 0.001, \eta_p^2 = 0.20) \), while the Treatment effect is not \( (F(2, 54) = 1.61, p = 0.21, \eta_p^2 = 0.06) \), but these must, of course, be interpreted in light of the interaction effect.\(^4\)

The participants’ mean test scores displayed in Figure 6.1 and Table 6.1.

\(^4\) Levene’s tests were not significant, meaning that homogeneity could be assumed; however, Mauchly’s test of sphericity was statistically significant; thus, the Greenhouse-Geisser correction was used. If this correction is not used, the Time-by-Treatment interaction effect is statistically significant \( (F(4, 108) = 2.555, p = 0.04, \eta_p^2 = 0.09) \).
What these results suggest is that, for a certain population of learners, feedback in the form of tree diagrams might, indeed, be more beneficial than less informative conditions of interpretation practice. Had data not been collected from L2 Japanese learners at American University (AU) and The George Washington University (GW), the relative effectiveness of the different treatment conditions might have seemed more straightforward. On the other hand, trends related to the set of research questions regarding the role of individual differences might have been less striking.

Figure 6.1. Comparison of treatment groups’ test scores, including only Georgetown and the University of Maryland
Table 6.1. Descriptive statistics of test scores across conditions, including only GU and UMD

<table>
<thead>
<tr>
<th>Test</th>
<th>Treatment Condition</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>No Feedback</td>
<td>59.42</td>
<td>10.44</td>
<td>54.27</td>
<td>64.56</td>
</tr>
<tr>
<td></td>
<td>Right/Wrong Feedback</td>
<td>62.45</td>
<td>8.10</td>
<td>57.57</td>
<td>67.34</td>
</tr>
<tr>
<td></td>
<td>Trees Feedback</td>
<td>58.64</td>
<td>13.56</td>
<td>53.62</td>
<td>63.65</td>
</tr>
<tr>
<td>Post-test</td>
<td>No Feedback</td>
<td>67.01</td>
<td>12.78</td>
<td>60.22</td>
<td>73.81</td>
</tr>
<tr>
<td></td>
<td>Right/Wrong Feedback</td>
<td>66.72</td>
<td>14.20</td>
<td>60.27</td>
<td>73.17</td>
</tr>
<tr>
<td></td>
<td>Trees Feedback</td>
<td>74.23</td>
<td>15.92</td>
<td>67.62</td>
<td>80.85</td>
</tr>
<tr>
<td>Delayed post-test</td>
<td>No Feedback</td>
<td>62.74</td>
<td>12.84</td>
<td>56.72</td>
<td>68.76</td>
</tr>
<tr>
<td></td>
<td>Right/Wrong Feedback</td>
<td>66.26</td>
<td>13.24</td>
<td>60.55</td>
<td>71.97</td>
</tr>
<tr>
<td></td>
<td>Trees Feedback</td>
<td>73.40</td>
<td>12.09</td>
<td>67.54</td>
<td>79.26</td>
</tr>
</tbody>
</table>

No Feedback n=18, Right/Wrong Feedback n=20, Trees Feedback n=19

From a more individualized perspective, looking again at the performance of all of the participants from all of the universities where the experiment was run, the parallel coordinate plots presented in Appendix N clarify that, whereas no learners in the No-Feedback or Right/Wrong conditions scored 100% on the tests, there were 2 participants in the Trees-Feedback condition who did so. Only 1 participant in the Right/Wrong condition, and none in the No-Feedback condition, scored above 90%. At a lower threshold of 80%, in comparison to
the 4/26 learners (15%) in the No-Feedback condition and 6/29 learners (21%) in the Right/Wrong condition who scored at least that high on the post-test, 10/25 learners (40%) in the Trees condition did so.

The differences in these tallies are not extreme enough for chi-square tests to indicate statistically significant differences among the groups. Nevertheless, from a more qualitative perspective, it may be telling that the only condition in which any learners achieved a perfect score was the Trees condition. As mentioned above, it was intentionally built into the design of the testing and treatment materials that, although rules of thumb based on somewhat less technical metalinguistic concepts (e.g., nominative versus dative case) might have allowed participants to show some improvement, it would not be possible to achieve a perfect score without developing sensitivity to c-command (or clause status) and an accurate sense of subjecthood, independent of superficial morphological markings. Furthermore, since some of the test items involved structures and case-markings that had never appeared during the treatments, the perfect scores in the Trees group cannot be attributed purely to memory for previously encountered sentence types; they necessarily entailed some amount of generalization.

So, is it possible for visual feedback in the form of tree diagrams to help learners to interpret sentences with reflexives more accurately? In the context of the present experiment, it would not be justified to claim that tree diagrams were more helpful in general than simple feedback informing learners about the correctness of their interpretations, or even that they were more helpful in general than not receiving feedback at all. However, it does seem quite reasonable to suggest that tree diagrams can be helpful for at least some learners (enabling
some in the Trees group, even, to achieve perfect scores on the post-test, and appearing to be more effective than the other conditions of practice when students at only certain universities are considered). This may be sufficient at least to call into question a claim that feedback should never be able to influence the learning of a target for which relevant distinctions are not represented phonologically.

As reviewed in Chapter 1, S. E. Carroll (2001, 2007) has predicted that certain areas of language should be impervious to the influence of feedback. This is based on assumptions she holds regarding the architecture of the language faculty, the nature of awareness, and the way in which she has concluded that learners must process feedback in order for it to lead to learning. In brief, assuming that learners’ awareness of language is mediated through phonological representations (Jackendoff, 1987, 2007), and assuming that awareness is necessary for a piece of feedback or metalinguistic information to function as negative evidence, Carroll claims that linguistic phenomena which are not available to awareness via phonological representations are inherently unable to be learned on the basis of feedback.

In Carroll’s Autonomous Induction Theory, she highlights not only that there is more to language than what is instantiated in surface forms, but also that for many linguistic phenomena, there simply are no concepts, metalinguistic or otherwise, that map onto them precisely (2001, 2011). These are valid and quite valuable insights which have not tended to be emphasized in empirical studies on the effectiveness of L2 feedback. Many researchers, from a range of theoretical perspectives, put forward variants of the claims that metalinguistic knowledge cannot be transformed directly into linguistic knowledge (e.g., Schwartz, 1999), that declarative knowledge cannot be transformed directly into procedural knowledge (e.g., Paradis,
and that explicit knowledge cannot be transformed directly into implicit knowledge (e.g., N. Ellis, 2005). Ideas about the modularity of the mind and the neurofunctional anatomy of different memory systems in the brain have led to some analogous conclusions. To supplement these rationales, Carroll emphasizes the importance of considering whether it is possible for the specific linguistic structures of interest to correspond to conceptual structures.

In relation to all of these perspectives, it is important to stress that the current study did not set out to demonstrate that conceptual representations can be converted directly into grammatical representations within the language faculty. Rather, drawing insights from N. Ellis’ (2005) perspective on the so-called ‘weak interface’ between explicit and implicit learning processes discussed in Chapter 1, one idea inspiring this project was that it might be possible to circumvent both a potential lack of direct correspondences between different types of mental representation and Carroll’s argument (valid or not) that learners cannot be aware of linguistic phenomena that are not mediated phonologically. Linguists have developed alternative formats to represent underlying structure, and tree diagrams can convey binding relationships visually as opposed to phonologically. If (some) learners can internalize this information in ways that enable them to repeatedly interpret L2 sentences more accurately, then it stands to reason that more targetlike linguistic knowledge might be developed in parallel with (and possibly due to guidance provided by) their uses of metalinguistic information.

It is also very important to point out, however, that although there may be reasons to believe that Carroll’s predicted constraint on the potential effectiveness of feedback is wrong, the findings of this research so far do not allow us to pinpoint which of Carroll’s assumptions underlying that prediction might be wrong. That is, they do not allow us to determine whether
it is because awareness can be mediated via non-phonological representations, whether it is because awareness is not required for feedback to be effective, or whether it is for some other reason entirely. Analyses of the introspective data which were gathered via concurrent think-aloud protocols, planned as the next step in this research project, could ultimately prove to be a crucial linchpin in the argumentation. Transparent links between information in the tree diagrams, verbalizations of reasoning processes making use of relevant metalinguistic concepts, and correspondingly more accurate interpretations of sentences on the post-test by participants in the Trees condition could provide support for the idea that the metalinguistic information made them aware of constraints governing interpretations of zibun and this facilitated improvement.

Still, as mentioned in Chapter 1, there have been some CALL feedback studies in which participants receiving only information regarding the correctness of their responses have shown levels of improvement similar to those of participants in more explicit instructional conditions (e.g., Camblor-Portilla, 2006; Moreno, 2007; Sanz & Morgan-Short, 2004). In relation to these findings, it has been argued, for example, that sometimes “locating the source of the problem and rejecting hypotheses is possible even without explicit feedback” (Sanz & Morgan-Short, 2004, p. 73). When the relevant features of linguistic targets are uncomplicated or familiar enough that learners already have pertinent metalinguistic knowledge at their disposal, simple right/wrong feedback might be considered equivalent, in a sense, to providing more detailed explicit information. In the context of the current study, with this caveat in mind, claims that the tree diagrams themselves led to more accurate sentence interpretations would be further supported by finding that reasoning processes involving concepts related to c-
command and subjecthood were substantially more likely in the Trees condition than in the others, and particularly among those learners in the Trees condition who showed improvement. The support would be even stronger if it were found that learners in the Trees condition who did not appear to know about c-command at first seemed to gain an incipient sense of it during the treatments, whereas this was much less likely in the other conditions.

In view of these concerns, it is worth reviewing what the linguistic targets have been in those studies where explicit feedback was not found to be more helpful. In Camblor-Portilla’s (2006) experiment, the target forms were gender- and number-agreement morphemes in Spanish, which, the author points out, involved straightforward and familiar concepts for the participants since noun-adjective agreement is taught very early in the curriculum (p. 74) and their L1, English, often marks plurality in the same way (by adding -s). The target in the studies by Sanz and Morgan-Short (2004) and Moreno (2007) was preverbal direct-object clitics. In the former study, the participants had to figure out the meanings of sentences (i.e., who was doing what to whom) based on the cases of the pronouns, which were limited to 4 different morphological forms. In the latter, they had to describe pictures, and the task simply involved choosing the correct pronouns to make sentences. In each of these studies, it makes sense that additional metalinguistic information might not have been necessary for the learners to understand the relevant distinctions.

In contrast, in those studies where advantages for explicit feedback have been found, the researchers have tended to argue that the targets were complex. The case particles focused on in Nagata’s (1993, 1995) studies are known to cause persistent problems for L1-English learners of Japanese since they entail subtle distinctions of meaning that do not exist in English
(e.g., using a different particle to express the location of a static activity versus a dynamic one versus one involving directed motion). Nagata did not find an advantage for more explicit feedback on simpler targets involving verb conjugations and vocabulary, but she did find one for the particles. In Lado’s (2008) study, which investigated learners’ ability to assign semantic roles to nouns in Latin sentences, the linguistic target and the tasks employed were similar to those in Sanz and Morgan-Short’s (2004) research; however, Lado describes her target as more complicated than Sanz and Morgan-Short’s in the sense that 8 novel morphological case-markings were involved and the learners had never before been exposed to Latin. In Rosa and Leow’s (2004b) experiment, where advantages for explicit information were also found, the linguistic target of past counterfactual conditionals in Spanish is semantically and formally quite complex, involving multiple time frames and verbal morphemes which must be carefully coordinated with each other across a biclausal structure: In the antecedent clause, the verb must be expressed in the pluperfect subjunctive, with the main verb in its past participial form and an auxiliary in the past subjunctive agreeing in person and number with the subject. In the consequent clause, the verb must be expressed in a conditional form agreeing in person and number with the subject.

In short, there seems to be a relatively clear trend emerging which suggests that future CALL feedback research may profit from including targets differing in complexity (and familiarity) in the same study. This is likely to pose challenges (DeKeyser, 2005). When de Graaff (1997), for instance, intended to include both simple and complex morphological and syntactic targets in his research into the effects of more versus less explicit instruction, the results led him to rethink his original judgments of complexity. In the design of his experiment, he considered

315
noun pluralization and the position of negative elements to represent simple targets, and
imperative formation and the position of topicalized pronominal objects to represent complex
ones. Reporting clear evidence of an advantage for explicit information, de Graaff states that he
found “no clear evidence” of a “differential effect of explicit instruction on the acquisition of
simple as opposed to complex structures” (p. 270). However, focusing solely on the syntactic
targets, de Graaff specifies that explicit instruction was, indeed, “much more effective” for the
complex syntactic structure than for the simple one, and he conjectures that the reason a
differential advantage for explicit instruction was not found on the complex versus simple
morphological structures as well may have been because the ‘simple’ morphological structure
was actually difficult: Participants in the explicit condition outperformed those in the implicit
condition on (supposedly simple) noun pluralization to a much greater extent than had been
expected. Since the difference between the plural endings was a nonmeaningful one which
depended exclusively on form (vowel harmony, which seems likely to have been unfamiliar),
explicit instruction on it “was at least as necessary and effective as for the meaningful
imperative morphology” which had been intended to represent a more complex learning target
(p. 270). In other words, de Graaff could perhaps have placed more emphasis on the finding
that explicit information was more helpful for a complex syntactic target.

Some speculation from Rebuschat and Williams (2009), though not based on a study of
L2 feedback in CALL, may also be germane to this trend. In seeking to explain why only a limited
amount of learning was observed in their experiment, they mention (among other possibilities)
that the participants’ existing metalinguistic knowledge may even have served as a distraction.
The linguistic target in their research was the placement of the verb phrase in relation to the
type of clause (main or subordinate) and the type of clause sequence (main-subordinate or subordinate-main) in a semi-artificial grammar using English sentences rearranged to follow certain characteristics of German syntax. Rebuschat and Williams explain that categories such as ‘subject’, ‘verb’ and ‘object’ may have been more readily available to the participants: “It could... be that categories such as clause type or clause sequence are simply not perceived to be relevant elements of grammar” (pp. 5-6). The target in the present experiment, as we have seen, involved ambiguous sentences and abstract metalinguistic concepts that L2 learners tend not to have learned about unless they have taken courses in theoretical linguistics.

Finally, it may be important to emphasize again that although researchers in the area of CALL feedback are undoubtedly interested in implicit learning processes, they do not tend to suggest that purely implicit learning is taking place in the less-explicit treatment conditions of their studies. The findings at this point may simply indicate that when participants across all treatment groups of an experiment are attempting to learn a complex target explicitly, more information can be helpful. Naturally, issues concerning the value of metalinguistic information in relation to the complexity of the linguistic targets may also be modulated by the influence of learners’ aptitudes.

Aptitude-treatment interactions (Research Questions 2 and 3)

The second research question driving the present experiment was whether individual learner differences in a variety of areas (i.e., visual memory, memory for linguistic material, knowledge of metalinguistic terms, sensitivity to ambiguity, grammatical sensitivity, enjoyment of grammar, and experience studying Japanese and linguistics) would show relationships with
learners’ accuracy in interpreting Japanese sentences containing the reflexive *zibun*. The third research question, exploring the issue of aptitude-treatment interactions, took the investigation further to ask whether any differences would be found in the relationships between IDs and *zibun* test performance according to treatment condition. After reviewing the overall relationships which emerged when all of the participants were considered together as a single group, the discussion will focus on each experimental group separately, and some provisional conclusions will be drawn regarding the abilities and background variables which turned out to be associated with success when learners were provided with different amounts and kinds of information. Tables 6.2 and 6.3 summarize which individual-difference measures showed statistically significant correlations with test scores in this experiment, both for all of the participants considered together (Table 6.2) and for subgroups of participants separated according to treatment condition (Table 6.3).

*Table 6.2. Statistically significant correlations between IDs and test scores (all participants)*

<table>
<thead>
<tr>
<th>Individual-Difference Measure</th>
<th>Pre-Test</th>
<th>Post-Test</th>
<th>Post-Test Controlling for Pre-Test</th>
<th>Delayed Post-Test</th>
<th>Delayed Post-Test Controlling for Pre-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPT (visual STM)</td>
<td></td>
<td>.30</td>
<td>.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT IV (gram. sensitivity)</td>
<td>.32</td>
<td>.35</td>
<td>.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT V (rote memory)</td>
<td></td>
<td>.30</td>
<td>.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT Total (short form)</td>
<td>.27</td>
<td>.28</td>
<td>.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reflexives Subtotal</td>
<td>.29</td>
<td>.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambiguity Subtotal</td>
<td>.23</td>
<td>.24</td>
<td>.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammar Subtotal</td>
<td></td>
<td>.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambig. &amp; Gram. Total</td>
<td>.28</td>
<td>.28</td>
<td>.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Japanese Study</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Course Level</td>
<td>.24</td>
<td>.41</td>
<td>.35</td>
<td>.24</td>
<td></td>
</tr>
<tr>
<td># Linguistics Courses</td>
<td>.25</td>
<td></td>
<td></td>
<td></td>
<td>.24</td>
</tr>
<tr>
<td>Enjoyment of Grammar</td>
<td>.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6.3. Statistically significant correlations between IDs and test scores (by treatment group)

<table>
<thead>
<tr>
<th>Individual-Difference Measure</th>
<th>Pre-Test</th>
<th>Post-Test</th>
<th>Post-Test Controlling for Pre-Test</th>
<th>Delayed Post-Test</th>
<th>Delayed Post-Test Controlling for Pre-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NO-FEEDBACK CONDITION</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VPT (visual STM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT IV (gram. sensitivity)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT V (rote memory)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT Total (short form)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reflexives Subtotal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambiguity Subtotal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammar Subtotal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambig. &amp; Gram. Total</td>
<td>.47</td>
<td>.43</td>
<td>-.43</td>
<td>-.49</td>
<td></td>
</tr>
<tr>
<td>Years of Japanese Study</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Course Level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Linguistics Courses</td>
<td>.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enjoyment of Grammar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RIGHT/WRONG-FEEDBACK CONDITION</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VPT (visual STM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT IV (gram. sensitivity)</td>
<td>.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT V (rote memory)</td>
<td></td>
<td>.47</td>
<td></td>
<td>.56</td>
<td></td>
</tr>
<tr>
<td>MLAT Total (short form)</td>
<td>.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reflexives Subtotal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambiguity Subtotal</td>
<td>.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammar Subtotal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambig. &amp; Gram. Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Japanese Study</td>
<td>.55</td>
<td>.53</td>
<td>.41</td>
<td>.45</td>
<td></td>
</tr>
<tr>
<td>Course Level</td>
<td>.56</td>
<td>.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Linguistics Courses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enjoyment of Grammar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TREES-FEEDBACK CONDITION</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VPT (visual STM)</td>
<td>.45</td>
<td>.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT IV (gram. sensitivity)</td>
<td>.56</td>
<td>.45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT V (rote memory)</td>
<td></td>
<td>.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLAT Total (short form)</td>
<td>.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reflexives Subtotal</td>
<td>.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambiguity Subtotal</td>
<td>.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammar Subtotal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambig. &amp; Gram. Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Japanese Study</td>
<td>.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Course Level</td>
<td>.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Linguistics Courses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enjoyment of Grammar</td>
<td>.48</td>
<td>.41</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What might be gleaned from Table 6.2 is that on the pre-test, overall, the participants’ scores were related to their grammatical sensitivity (MLAT IV), sensitivity to ambiguity in English (perhaps particularly in sentences containing reflexives), Japanese course levels, numbers of linguistics courses taken, and enjoyment of grammar. On the post-test, all of these IDs except for the number of linguistics courses the participants had taken and their enjoyment of grammar still showed statistically significant relationships with performance. Additionally, after the learners had engaged in the treatment activities, their test scores were statistically significantly related to their visual short-term memory (VPT), rote memory for linguistic material (MLAT V), and knowledge of metalinguistic terminology. By the time of the delayed post-test, however, the only ID variable showing a statistically significant relationship with test scores was the participants’ Japanese course levels, at the same strength as had been found on the pre-test. The only ID not to show a relationship with test performance overall was the length of time the learners had been studying Japanese.

When pre-test scores were controlled for, nearly all of the same IDs showed statistically significant partial correlations with post-test scores as when correlations had been run with post-test scores alone—the exceptions being the participants’ knowledge of metalinguistic terminology and their sensitivity to ambiguity in sentences with reflexives in English. For the delayed post-test, whereas canonical correlations had produced a statistically significant relationship only with the participants’ Japanese course levels when pre-test scores were not controlled for, correlations partialling out the participants’ pre-test scores produced a statistically significant relationship only with the numbers of linguistics courses they had taken.
Intuitively, several of these findings seem to make sense. From the very outset of the study all the way through to the end, regardless of treatment condition, the activities involved reading complex sentences in a foreign language and deciding what they could and could not mean. It is therefore understandable that the participants’ Japanese course levels would emerge as a statistically significant predictor of performance on the pre-test, post-test, and delayed post-test. In reflecting on this result, it seems important to keep in mind that many of the target sentences involved complex structures (e.g., relative clauses, causatives) which may have been difficult for learners at lower proficiency levels to process even though they had been introduced in their Japanese language classes. A stronger foundation in Japanese could also possibly have made some participants’ *kanji* recognition more automatic, even for lexical items that all of the learners were expected to be familiar with, thereby making their L2 processing smoother. It also seems reasonable to consider that learners at higher levels of study would simply have developed intuitions closer to those of native speakers of Japanese, whereas learners in lower-level courses might have been influenced to a greater degree by transfer from English (if transfer was, in fact, an issue).

The participants’ grammatical sensitivity is another individual difference which, understandably, emerged as a significant predictor of performance on both the pre-test and the post-test. The MLAT Part IV measures sensitivity to the functions of words and phrases in English sentences—an ability that seems likely to contribute to learners’ capacity for processing complex L2 sentences. In the target sentences of this experiment, not only were there multiple types of clauses with different functions (e.g., relative clauses being adjectival, other embedded clauses functioning nominally in a sense), but there were also many non-canonical and
somewhat uncommon case-markings (e.g., nominative objects, dative subjects), with multiple role players mentioned in each sentence. On top of that, the sentences were not merely complex; they were potentially ambiguous. Accordingly, it makes sense that the participants’ sensitivity to ambiguity emerged as a predictor across the pre- and post-tests. Despite the design of the tasks, which sought to highlight only one possible meaning at a time, the learners’ ability to interpret the target sentences may have involved being receptive to or aware of the existence of multiple meanings, whether because they needed to play up weak readings or inhibit strong alternatives. This capacity was measured in English so that it would be influenced only minimally, if at all, by L2 Japanese proficiency, but it seems to be related to the interpretation of potentially ambiguous sentences in the L2 as well.

Regarding the ID measures that did not show relationships with pre-test scores, it makes sense that the participants’ rote memory (MLAT V) and visual short-term memory (VPT) would not yet be strong predictors of performance, considering that interpretations of zibun at the time of the pre-test could involve neither remembering visual diagrams of sentences nor remembering which types of sentences had been associated with positive or negative feedback. A bit more tenuously, an argument might also be made for why the participants’ knowledge of metalinguistic terminology was not very reliably related to their pre-test scores overall. It will be recalled that before taking the pre-test, the learners had just completed the test of sensitivity to ambiguity in English. On that test, they had been asked to generate paraphrases of ambiguous and unambiguous sentences, and their high levels of English competence may have produced natural intuitions about which they could feel fairly confident. Unless a participant had taken linguistics courses, in which structurally ambiguous sentences are often diagrammed
as a way of demonstrating why linguists postulate underlying structure, s/he may not have realized that ambiguities can be systematic. Furthermore, as the pre-test represented the first time that the participants were being asked to judge Japanese sentences of this nature, they were perhaps unlikely by that point to have developed a metalinguistic approach to it as a learning problem.

Interestingly, the participants’ experience with linguistics courses and their enjoyment of grammar emerged as overall predictors of pre-test performance without showing as reliable a relationship with post-test performance. Here, speculating is difficult. What can be noted, however, is that the only individual differences which produced statistically significant positive correlations with the number of linguistics courses the participants had taken were their sensitivity to ambiguity in English sentences ($\rho=.25$), including sentences with reflexives ($\rho=.30$), and their total scores on the test of sensitivity to ambiguity and knowledge of metalinguistic terminology ($\rho=.24$) (see Table 5.12). Moreover, the only ID which showed a statistically significant positive correlation with the participants’ enjoyment of grammar was their knowledge of metalinguistic terminology ($\rho=.23$). It seems to be the sensitivity itself, and the knowledge itself, which carried through to the post-test, as opposed to the participants’ background experience with linguistics or grammar enjoyment.

Moving on to the post-test, as alluded to above, it makes sense that positive relationships between test performance and the participants’ visual short-term memory, rote memory for language, and knowledge of metalinguistic terminology might emerge following the treatments. On the post-test, since some of the learners had received visual feedback, it seems reasonable to expect a relationship with visual short-term memory (and, indeed, to
anticipate the ATI results, it was the Trees-Feedback group who contributed the most to this positive relationship). Since some of the learners had received feedback that stood to be remembered or not, it also seems reasonable to expect relationships with rote memory for linguistic material (and, indeed, it was the Trees-Feedback and Right/Wrong-Feedback groups who drove these relationships). Finally, since at least some improvement could possibly have been achieved through the application of metalinguistic concepts, it seems reasonable to expect a relationship with knowledge of grammar (and, although different ATI results could also have been expected, this one seems attributable to the No-Feedback group, who had been left to their own devices).

The fact that none of the individual differences except the participants’ Japanese course levels and numbers of linguistics courses showed statistically significant relationships with delayed post-test scores seems less straightforward to explain. One might be tempted to argue that this could be related to the No-Feedback and Right/Wrong-Feedback groups’ lack of sustained improvement, but the ATI results reveal that it was in precisely those groups where some relationships were found. Another possibility that could be pursued is the idea that negative correlations in one group might have counteracted positive correlations in another. This may be valid where sensitivity to ambiguity is concerned since the learners in the No-Feedback condition showed an inverse relationship between their sensitivity to ambiguity and their delayed post-test scores. However, not all of the ID measures followed that sort of pattern.

What these last points highlight is that contemplating such general relationships on their own does not reveal nearly as much as examining relationships within each experimental group separately and then comparing those results with each other (in other words, exploring
aptitude-treatment interactions). It bears mentioning again that, with so many correlations run, at least some can be expected to be by chance. At the same time, given that fewer participants were included in each separate-group analysis, the correlations had to be stronger in order to produce statistically significant results. In short, as a means of uncovering themes that might be pursued further in subsequent analyses and experiments, a group-by-group exploration of relationships between IDs and test scores, tempered by appropriate caution, can be fruitful.

**No-Feedback condition**

In the No-Feedback condition, there were no statistically significant relationships between individual differences and pre-test scores. On the post-test, however, after the participants had engaged in practice interpreting sentences without receiving any feedback on the accuracy of their responses, three ID variables were strongly related to performance: grammatical sensitivity, knowledge of metalinguistic terminology, and the numbers of linguistics courses taken. The partial correlations with post-test scores controlling for pre-test scores mostly showed the same pattern despite not quite reaching statistical significance; those that approached significance (grammatical sensitivity, knowledge of metalinguistic terminology, and the total score on the test of ambiguity and metalinguistic terminology) were the same as had been found with straight post-test scores. For the linguistics-courses variable, differences between parametric and non-parametric statistics (i.e., partial correlations versus Spearman’s

---

42 This is not problematic in and of itself, but as we will see, the treatment groups showed different relationships between IDs and pre-test scores, which is unfortunate because it makes cross-group interpretations of the relationships between IDs and post-test scores more difficult.

43 Since the relationship between post-test scores and scores on the subtest of knowledge of metalinguistic terminology ($\rho=.47$) may be the major contributor to the relationship between post-test scores and total scores on the test of sensitivity to ambiguity and knowledge of metalinguistic terminology ($\rho=.41$), the latter relationship is not considered to represent an additional ID variable beyond those listed.
rho) may have played a role in the failure of that relationship to reach statistical significance considering the extreme positive skew in the data’s distribution and the fact, therefore, that an important assumption underlying partial correlations was violated.

An initial reaction to these results might be to conjecture that participants in the No-Feedback condition had taken a metalinguistic approach, and some of them (with higher grammatical sensitivity, knowledge of grammar, and linguistics experience) had been able to come up with rules of thumb that were at least partially correct. For example, perhaps some had hypothesized that zibun had to refer to the main subject of the sentence, or perhaps they had imagined that zibun could not refer to an object, but had an incomplete understanding of what counted as an object (e.g., misinterpreting the embedded subjects in causative structures as objects). In any case, since any such rules of thumb would never have been called into question by feedback to the contrary, these participants may have continued to develop and use their provisional rules, and the learners who came up with more accurate ones may have performed better on the post-test. This speculation will, of course, need to be verified by inspecting the think-aloud data in the next phase of this research project. It cannot simply be taken at face value here.

The strong negative relationship between the No-Feedback participants’ accuracy in interpreting ambiguous English sentences and their delayed post-test accuracy in interpreting ambiguous Japanese sentences with zibun (including when pre-test scores are controlled for) is also potentially very interesting. One additional finding which might be contemplated in relation to this is that the participants’ Japanese course levels were just shy of showing a similarly strong statistically significant relationship with their delayed post-test scores (\(\rho=.40,\))
If we keep in mind that the No-Feedback group had appeared to show pre-post gains which were no longer in evidence on the delayed post-test, and if we speculate based on this that they may have forgotten some of the rules that had (hypothetically) been helping them on the post-test, perhaps it could also be hypothesized that those participants with stronger intuitions about English sentences reverted to what felt natural. That is, perhaps a strength in Japanese (represented by course level) worked to the learners’ benefit, whereas a strength in English (represented by ambiguity scores) worked to their detriment, particularly given that so many of the target sentences had been designed to exemplify differences between the two languages in the availability of certain interpretations.

**Right/Wrong-Feedback condition**

The pattern of results in the Right/Wrong-Feedback condition puts us unavoidably face to face with the problem of differing relationships between IDs and pre-test performance among the three experimental groups. Despite the fact that there had not yet been any manipulated differences among the conditions on the first day of the experiment, the participants in the Right/Wrong group showed statistically significant correlations between their pre-test scores and their grammatical sensitivity and sensitivity to ambiguity in English, whereas this was not the case for the No-Feedback and Trees-Feedback groups. In striking contrast, the only ID variables showing statistically significant correlations with post-test scores in the Right/Wrong group were the number of years the participants had been studying Japanese and their Japanese course levels. The fact that it is precisely those IDs (years of study and course level) which showed statistically significant partial correlations with post-test scores
when pre-test scores were controlled for might increase our confidence in the validity of these results, as might the fact that both of those same IDs showed strong, statistically significant positive correlations with pre-post gain scores (years of study: \( \rho = .57 \), course level: \( \rho = .48 \)). However, confidence is not an explanation.

In attempting to make sense of these findings indirectly, one line of attack would be to contemplate why it might be that certain IDs unexpectedly did not show relationships with post-test scores in the Right/Wrong group. For example, it would have been understandable if the learners’ grammatical sensitivity (MLAT IV), knowledge of metalinguistic terminology, and experience with linguistics had produced significant correlations, just as they had in the No-Feedback condition, for similar reasons. In trying to resolve this discrepancy, we can refer to Figure 5.4, reproduced here as Figure 6.2.

*Figure 6.2. Scatterplots of relationships between MLAT IV and pre-/post-test scores, highlighting the Right/Wrong-Feedback condition*
It will be remembered from Chapter 5 that when subsets of learners in the Right/Wrong condition were examined (impressionistically), it seemed that participants with roughly mid-range scores on MLAT IV may have performed better on the post-test compared to the pre-test, whereas participants with higher scores on MLAT IV may have performed worse. Applying similar reasoning to that considered above for the No-Feedback group, one might speculate that learners with high grammatical sensitivity in the Right/Wrong-Feedback condition also came up with partially correct rules of thumb during the treatment sessions. However, since participants in the Right/Wrong-Feedback condition received feedback, these partially correct rules were disconfirmed and therefore abandoned by the time of the post-test.

This calls to mind Reber’s (1989) point that explicit approaches can be disadvantageous compared to implicit ones when learners try to search for complex rules that they cannot find. It also recalls some of Robinson’s (1997b) empirical results (discussed in Chapter 2), such as the finding that grammatical sensitivity was related to the learning of an ‘easy’ rule in a rule-search condition, whereas memory was related to the learning of a ‘hard’ rule in that condition. In an instructed condition, in contrast, both grammatical sensitivity and memory were related to learning both the easy rule and the hard rule. In the current study, perhaps the participants with mid-range grammatical sensitivity who seemed to show improvement took an approach more amenable to implicit learning. Alternatively, perhaps they came up with partially correct rules of thumb and simply did not realize upon getting feedback that the rules had been contradicted. Whatever the explanation, it seems clear that researchers must be attuned to the possibility of non-linear relationships.
Given this hypothetical account of the lack of a straight linear relationship between grammatical sensitivity and post-test scores in the Right/Wrong condition, it is also possible and important to reflect on why the participants’ years of Japanese study and Japanese course levels showed positive relationships with post-test performance. Japanese language instructors, after all, do not provide rules specifically related to interpretations of *zibun*. Assuming an explicit approach to the treatment activities, perhaps a strong foundation in the language (irrespective of grammatical sensitivity) put learners in a better position to recognize sentence structures they had been taught, to understand which types of sentences they were interpreting correctly versus incorrectly, to encode that information with higher fidelity, and to remember it better. Assuming a more implicit, memory-based approach, it seems likely that more fluid and accurate L2 processing could be associated with greater levels of success when being informed only of the targetlikeness of interpretations. If a learner’s processing of sentences has been piecemeal and choppy, it stands to reason that feedback might not be as clearly or directly linked to each sentence as a whole. Moreover, a richer framework of previous experience with the Japanese language might mean more places in the scaffolding where incremental new bits of information could fit in and catch hold, so to speak. In any event, participants’ experience with Japanese was associated with higher scores following treatment all the way through to the delayed post-test, at which point their years of Japanese study still showed a statistically significant positive relationship with test performance.

The other ID variable that emerged as statistically significantly related to delayed post-test performance in the Right/Wrong condition (both with and without pre-test scores controlled for) was the participants’ memory for linguistic material (MLAT V). This makes sense
considering that the learners in this group had been presented with information about their interpretations of sentences which could be remembered a week after the final treatment.

**Trees-Feedback condition**

Finally, in the Trees-Feedback condition, among several individual-difference variables which had shown relationships with pre-test performance (i.e., the participants’ sensitivity to the ambiguity of English reflexives, years of Japanese study, Japanese course levels, and enjoyment of grammar), none except enjoyment of grammar was then related to post-test scores. The IDs which did produce statistically significant positive correlations with post-test performance were visual short-term memory, grammatical sensitivity, rote memory for linguistic material, sensitivity to ambiguity in English, and (as mentioned) grammar enjoyment.

The latter set of relationships seems possible to explain straightforwardly: During the treatments, the Trees group had been provided with visual diagrams of sentence structure as feedback; thus, visual memory (VPT) was related to post-test performance. This visual feedback had depicted the grammatical functions of words in an abstract and possibly unfamiliar way; thus, grammatical sensitivity to the functions of words in sentences (MLAT IV) was related to post-test performance. The feedback on linguistic forms and their interpretations had stood to be remembered or not; thus, rote associational memory for linguistic material (MLAT V) was related to post-test performance. The interpretation activities with zibun had involved being sensitive to what sentences could and could not mean in Japanese; thus, accuracy in interpreting ambiguous and unambiguous sentences (on the test of sensitivity to ambiguity in English) was related to post-test performance.
Anecdotally, a few participants in the Trees-Feedback condition mentioned during the final debriefing at the end of the experiment that, hating grammar and failing to understand the diagrams, they had simply stopped paying attention to them. One participant (P74) good-naturedly explained that she had simply begun skipping the slides showing the tree diagrams; another (P83) expressed vehement opposition to them as a means of teaching or learning anything about Japanese; and another (P65) referred to them with a sort of bewildered amusement as “the weird spider diagrams.” In contrast, among the participants who enjoyed them, one (P22) mentioned, despite never having diagrammed sentences or having taken a linguistics class before, that understanding the diagrams had felt intuitive and natural. Another (P40) said that he felt he had learned more about Japanese grammar from the tree diagrams presented in this experiment than he had in his Japanese classes all semester (and this was during the final-exam period). In view of this range of reactions to the tree diagrams, it makes sense that enjoyment of grammar would have continued to be related to test performance following the treatments in the Trees condition.

It is somewhat less straightforward to explain why the Trees-group participants’ years of Japanese study and Japanese course levels would be related to their pre-test scores, but not to their post-test scores. (In fact, years of study showed statistically significant negative correlations with pre-post gains [\(\rho = -.44\)] and pre-delay gains [\(\rho = -.54\)], and course level showed a negative correlation with pre-delay gains that nearly reached statistical significance [\(\rho = -.38, p = .06\).] In seeking to interpret these results, we can take the same approach as that pursued with MLAT IV scores in the Right/Wrong-Feedback condition. Scatterplots from Appendices P5-P6 are reproduced here as Figures 6.3 and 6.4.
In Figure 6.3, which displays the relationships between years of study and test scores, red boxes are drawn around subsets of participants in the Trees group: those with under 4 years of Japanese study, those with between 4-6 years, and those with more than 6 years. Inspecting the data points in the under-4 range, it appears that everyone may have demonstrated at least some pre-post improvement, and several achieved relatively high scores on the post-test (if we consider a threshold of 80% to represent high scores). Among the participants with 4-6 years of study, it seems that most tended to improve, and some achieved post-test scores of 80% or above. For the participants who had been studying Japanese for more than 6 years, however, the data points seem to move down from pre- to post-test. One participant in that range, whose data point is circled, achieved a post-test score of 100%, but none of the other participants with more than 6 years of Japanese study achieved scores higher
than 80%. Also worth noting is that several participants with fewer than 4 years of Japanese study had more room for improvement, as compared to participants with a greater number of years of study who had performed better on the pre-test.

*Figure 6.4. Scatterplots of relationships between course levels and pre-/post-test scores, highlighting the Trees-Feedback condition*

In Figure 6.4, which displays relationships between the participants’ course levels and test scores, red boxes are drawn around subsets of Trees-group learners in their universities’ 4th- and 5th-semester Japanese courses, 6th- and 7th-semester Japanese courses, and beyond. All three subgroups seem to exhibit some evidence of improvement—with wide ranges, however, in each. There is some suggestion of pre-post losses in the highest course-level subgroup, but
also a perfect post-test score of 100% in that subgroup and certainly not enough data points to make any generalizations.44

In making sense of these patterns, one issue to keep in mind is that the participants’ lengths of Japanese study do not necessarily correspond to their Japanese course levels, and there may be reasons for this. For instance, some learners with very few years of study may have been able to place into particular university course levels in a shorter amount of time than others because they were talented language learners. Learners who had taken longer to reach the same course levels may also have had trouble achieving learning gains quickly over the course of this experiment. Alternatively, acknowledging that, from the lowest to the highest Japanese course levels, no one had been taught about the properties of zibun, one could reason that when learners are provided for the first time with explicit metalinguistic information regarding how an aspect of the L2 works, the amount of experience they have with the language does not matter as much as the amount of ability they have to make use of the metalinguistic information provided. This idea could be pursued further by generating scatterplots with each participant’s data point labeled so that it would be possible to trace specific people’s improvement or lack thereof in relation to their years of study, course levels, and grammatical sensitivity scores.

At the very least, a way of double-checking the intuitive validity of these analyses is by imagining that the results had been reversed and trying to explain those imagined outcomes.

For example, we can imagine that the Right/Wrong-Feedback group had shown relationships

44 Another important consideration is that, compared to the Right/Wrong- and No-Feedback groups, the Trees group had a greater number of participants who had been studying for more than 6 years and who had reached the highest course levels; thus, there were more data points in the higher ranges of those variables for the Trees group. This is important to keep in mind when making comparisons of correlation strengths across the treatment conditions of this study.
between the participants’ post-test scores and their visual short-term memory and grammatical sensitivity, whereas the Trees-Feedback group had shown relationships only between the participants’ post-test scores and their years of Japanese study and Japanese course levels. A pattern like that seems potentially more difficult to explain.

Beyond these approaches, a more conservative policy would be to admit speculation regarding only those individual differences which showed relationships with post-test scores and delayed post-test scores when pre-test scores were partialled out (keeping in mind, however, that the assumptions for parametric statistics were not met and that the power to find relationships might therefore have been reduced). Under that policy, the following patterns remain relevant: In the No-Feedback condition, the only statistically significant relationship was an inverse one between the participants’ sensitivity to ambiguity in English and their scores on the delayed post-test. In the Right/Wrong-Feedback condition, the variables pertinent to post-test performance were the participants’ years of Japanese study and Japanese course levels, while the variables pertinent to delayed post-test performance were years of Japanese study and rote memory for linguistic material. In the Trees-Feedback condition, the important variables seemed to be visual short-term memory and grammatical sensitivity.

These results could possibly be interpreted as suggesting that when participants were not given feedback (in the No-Feedback condition), those with clearer or stronger intuitions about the meanings of English sentences performed worse when presented with Japanese sentences whose meanings were designed to differ from their superficial translation equivalents in English. When participants were informed about the correctness of their interpretations without being given additional information (in the Right/Wrong-Feedback
condition), those with greater memory were able to recall it at least a week after the treatments had ended, and those with more Japanese experience may have had richer frameworks for internalizing the feedback in the first place. When participants were informed about the correctness of their responses and given additional information in the form of visual diagrams (in the Trees-Feedback condition), those with stronger visual memory and capacity for understanding the functions of words in sentences were better able to internalize and make use of feedback.

This more limited set of interpretations might ignore some of the interesting trends in the data, but what it does nonetheless indicate quite unmistakably is that merely comparing test scores across the different treatment conditions (as in Research Question 1) does not provide nearly a complete picture of what happened in this study, and examining relationships between individual differences and test scores for all of the participants combined (as in Research Question 2) does not provide nearly a complete picture of what happened. Undoubtedly, the analyses presented just above in relation to Research Question 3 do not succeed in providing a complete picture, either, but they do represent a step in the desired direction which can be supported and embellished (or refuted) by consulting additional sources of information, such as introspective data from the reflection questionnaires and think-aloud protocols. In any event, the results so far substantiate Robinson’s (2002) argument that the “true nature” of the influence of individual differences on learning will be found in interactions, and “this is not, ultimately, reducible to general statements that IDs do, or do not, affect global incidental, explicit, or implicit learning” (p. 260).
Limitations and future research

The limitations of this research derive from several sources: practical, methodological, and theoretical. One immediately obvious concern is the fact that the three treatment sessions were spread over so many days (10 on average). Moreover, there was variation in the time that elapsed from treatment to treatment for different participants: Some took roughly a week and a half to complete all three, whereas others took more than 2 weeks. Even though this problem affected the treatment conditions equally (see Chapter 4), it is certainly less than ideal considering the amount of memory decay that may have occurred between sessions and the amount of noise it added to the data. Unfortunately, it was more or less unavoidable given that the research participants were unpaid volunteers with numerous other responsibilities, from homework, midterms, and final papers to off-campus jobs, orchestral rehearsals, and sports practice. Further analyses could seek to determine whether the amount of time that elapsed from pre- to post-test had any bearing on test performance. Future experiments could be improved by providing greater incentives to participants to show up on a tighter schedule (e.g., course credit and money, as opposed to fruit and candy). One of the universities’ Japanese programs had a strict prohibition against offering extra credit for their students to participate in research; thus, this option was not pursued at any of the universities. Regarding funding, several grant applications were submitted before data collection began, and one was successful: The journal Language Learning awarded a generous Dissertation Grant of $2,000; however, those funds were designated for the purpose of paying a bilingual research assistant to transcribe the think-aloud protocols, a process which has begun and will be completed once additional funds can be gathered.
Another source of noise in the data is the fact that the experiment took place not only over multiple semesters, but also over the span of each individual semester, meaning that some participants may have been mentally fresh throughout their participation, whereas others reported feeling exhausted and experiencing academic stress related to midterms or finals.

Data collection also took place in a variety of research locations. These ranged from a bright, clean, spacious, technologically sophisticated computer laboratory to a small, rather dimly lit and cramped space with much older computers. It is less obvious how environmental differences may have impacted the results, but attrition problems, at least, seemed to be somewhat more severe among participants whose schedule openings matched only the times when the small, dimly lit lab was available.

A methodological issue which must be addressed in future research is the similarity of the testing and treatment materials. Based on previous research into L2 learners’ interpretations of reflexives, the most valid task type to use appeared to be a picture- and story-based truth-value judgment task (see Chapter 4 and Appendix C). As such, this highly controlled task type was employed for all activities of the study. However, that being the case, a major caveat is that the results cannot be taken to imply anything about how the participants might have used or interpreted zibun in more naturalistic communicative situations or even in more authentic and contextualized reading (or writing) situations. Facing a similar situation, Carroll and Swain (1993) point out that their experimental context was different from contexts involving spontaneous speech in that “the semantic information and the syntactic forms... were predetermined. Thus, the decision task... [was] greatly simplified in comparison to the computational demands made of learners in ‘naturalistic’ contexts” (p. 372).
A complication in attempting to design other types of tasks to evaluate learners’ interpretations of sentences containing reflexives is that it is necessary to pin down a clear sense of what learners think sentences both can and cannot mean when some of the sentences are ambiguous and others are not. Multiple-choice tests of learners’ (and native speakers’) interpretations of reflexives (e.g., Hirakawa, 1990; Matsumura, 1994; Thomas, 1989, 1991; Yuan, 1998) have been argued to suffer from the problem of preferred readings, and simply asking learners to generate possible readings seems even less likely to work, considering how well the test of sensitivity to ambiguity in English functioned as an individual-difference measure in this study despite the fact that multiple interpretations for ambiguous sentences are theoretically part of any native speaker’s competence. Furthermore, since many of the sentence structures of interest in this area are quite complex, avoidance might be an issue in tests of production which attempt to elicit them. If *kare-zisin* (‘he-self’, a locally bound anaphor) could be included as an alternative to *zibun*, a multiple-choice or fill-in-the-blank format might be attempted, but preferences might play a role there as well, and the treatment sessions might need to be longer in order to include enough exemplars.

None of this is to say that alternative methods are impossible to imagine; an area of growing interest in the field of SLA, especially where learners’ interpretations of sentences are concerned, is in the opportunities provided by eye-tracking software to explore which words or morphemes learners focus on and regress to in L2 sentence processing. Such data might be difficult to interpret, but it could be fascinating to explore learners’ eye movements on ambiguous versus unambiguous sentences, perhaps particularly in cases where negative evidence would theoretically be required for them to reject an antecedent allowed in their L1.
It is not only in relation to transfer of training that questions related to task type can be illuminating. As discussed in Chapter 2, researchers have found different patterns of results on different types of outcome measures, and even regarding which individual-difference variables were associated with stronger performance on each outcome measure (e.g., Erlam, 2005; Lado, 2008; Révész, to appear (b); Robinson, 2002). The study of aptitude/treatment/test-type interactions, like the study of aptitude/treatment/learning-target interactions mentioned by DeKeyser (2003), can “provide much more insight into all three of these factors than the study of any one of them in isolation can hope to accomplish” (p. 337).

A limitation which has already been described in the discussion of RQ3 is the fact that there were differences among the treatment conditions with regard to the relationships shown between individual differences and pre-test scores. For example, in the Right/Wrong-Feedback group, there was a statistically significant relationship between grammatical sensitivity and pre-test scores which did not appear in the other groups and which was no longer in evidence on the post-test, whereas in the Trees-Feedback group, there was a statistically significant relationship between years of Japanese study and pre-test scores which did not appear in the other groups and which was no longer in evidence on the post-test. This somewhat hindered the ability to draw crystal-clear conclusions about the relationships between IDs and post-test scores. Ideally, the treatment groups would have shown the same relationships between IDs and pre-test scores, and then different relationships between IDs and post-test scores.

Collecting data from a greater number of participants might resolve this problem. Another possibility, although it might be inadvisable due to violations of statistical assumptions, would be to run repeated measures analyses of covariance (RM ANCOVAs) on the data,
performing a separate ANCOVA for each individual-difference variable. ANCOVAs are said to provide a ‘purer’ measure of treatment effects by holding a certain variable (the covariate) constant and thereby reducing the amount of unexplained variance in the model or eliminating a possible confound (Larson-Hall, 2010). Field (2009) describes ANCOVA as being similar to a hierarchical regression with the covariate entered in the first block and the experimental manipulation entered in the second block. If the effect of the covariate is statistically significant, it means that that variable had an independent effect on how the participants performed on the dependent measure. If the treatment effect is statistically significant when the covariate is included, it means that the manipulation had an effect which is possible to see when the covariate is controlled for. If not, it means that treatment group does not explain variance in the model. In the context of the current experiment, particularly of interest would be the Time-by-Treatment-by-Covariate interaction for each individual difference measure. A statistically significant effect of that nature could indicate that the ID in focus was associated with improvement over time differentially according to treatment condition. A problem, however, is that the ID data in this study often violated the assumptions for parametric statistics, meaning that RM ANCOVAs might not be appropriate.

One further concern is that although data were gathered on the participants’ years of Japanese study and Japanese course levels, their general proficiency was never measured directly. Unfortunately, it would be ill-advised to use either of the former as a proxy for the latter. According to Thomas (1994), the fact that institutional standards differ “especially threatens the validity of research that pools learners holding a particular status in one institution with learners holding the same status in another institution” (p. 317). In the present
experiment, the participants were randomly assigned to different treatment conditions, and pre- and post-tests were employed to investigate learning gains in a narrowly specified area of language. However, for the correlational analyses involving years of study and course levels, it is important to bear in mind the wide variability in what those could mean for different participants. A year of studying Japanese twice a week in middle school, for instance, may be very different from a year of studying Japanese every day in high school, and a fifth-semester course at one university, for example, may be very different from a fifth-semester course at another university. No statistically significant differences were found among the treatment groups with regard to the participants’ years of study or course levels, but this does not necessarily mean that there were not Japanese proficiency differences.

While not exactly a limitation, it is also worth mentioning that the data from many other individual-difference measures remain unanalyzed. As explained in Chapter 4, the participants completed several motivation questionnaires over the course of the experiment as well as a reflection questionnaire at the end (Appendices G1-G5). They also took operation-span and listening-span tests of their working memory capacities (Appendices H1-H2), and they produced confidence judgments and source attributions for all of the test items (Appendix D1). Any or all of these sources of information could lead to more detailed and accurate interpretations of the results of this experiment than have been presented thus far.

Regarding the issue of motivation, for example, it could be explored whether learners who received less explicit feedback (or none at all) might have reported exerting less effort than those who received more explicit feedback, and whether the amount of effort they reportedly expended had any independent relationship with test performance. If such
associations were found, they would be necessary to acknowledge when reasoning about the apparent benefits of providing metalinguistic information because they would suggest that the value of having been assigned to the most explicit treatment condition might not be solely attributable to the additional informational content per se. It would also be interesting to know whether particular aspects of aptitude might have been associated with particular motivational characteristics and whether those relationships differed at all across the experimental groups. It is certainly possible to imagine, as just one hypothetical example, that learners with low visuo-spatial abilities might have felt frustrated in the Trees-Feedback condition and ultimately made less of an effort, whereas learners with similarly low visuo-spatial abilities in the Right/Wrong-Feedback condition might have maintained higher levels of motivation and continued to make an effort since they were spared the frustration of dealing with visually oriented feedback. Some inter-relationships among treatment conditions, aptitudes, and motivation might be robust enough to emerge in statistical analyses, whereas other quite intricate interactions might not, in which event a case-study approach might be revealing.

By analyzing the confidence ratings and source attributions produced by the participants on each test, it might also be possible to draw some inferences concerning the types of knowledge that learners (and learners with particular profiles of abilities) may have been developing in each condition. This would also help to refine conclusions regarding whether or not there was a treatment effect in each condition. One line of speculation might be to hypothesize that perhaps the Trees-Feedback condition was effective at promoting explicit rule-based knowledge which was immediately applicable to multiple test items and allowed pre- to post-test improvement to be identified statistically over the relatively short duration of the
study, whereas the Right/Wrong condition was effective at promoting intuitions which were developing incrementally and might have taken longer to show up statistically in pre-post gains.

In the literature on artificial grammar learning (e.g., Dienes, Altmann, Kwan, & Goode, 1995; Dienes & Scott, 2005; Rebuschat & Williams, 2009), relationships between learners’ levels of confidence and accuracy are interpreted as indicating what types of knowledge they may have developed: If they show a positive relationship between confidence and accuracy, for instance, they are considered to have developed conscious judgment knowledge. If they report no confidence, but nonetheless perform accurately on certain items, they are considered to have developed some unconscious judgment knowledge. It will be interesting to compare this experiment’s No-Feedback, Right/Wrong-Feedback, and Trees-Feedback conditions with respect to the proportions of judgments (and accurate judgments) reportedly based on guesses, intuition, memory, and rules, as well as the participants’ levels of confidence when making each type of judgment, both for accurate and inaccurate responses. Considering the existence of aptitude-treatment interactions, it will also be interesting to examine these issues in relation to factors such as grammatical sensitivity and knowledge of metalinguistic terminology. It seems reasonable to expect, for instance, that a hypothesis that the Trees-Feedback condition promoted explicit rule-based knowledge might seem to be disconfirmed until qualified with the phrase ‘…among learners with high grammatical sensitivity.’

Then, of course, there are the data from the think-aloud protocols, which will allow for analyses of what the learners were aware of across the treatment conditions, how certain types of awareness reports may have been related to test performance both overall and across the conditions, and whether reports of awareness were related to individual learner variables. The
participants were given the choice of speaking in the L1 and/or the L2 due to serious concerns about potential non-veridicality and reactivity that would have arisen had they been limited to speaking in one language or the other. However, since I do not speak Japanese, I am unfortunately unable to transcribe most of the think-alouds myself. With over 100 hours of verbalization data to be transcribed, additional funding must be sought to complete the process.

In order to move beyond drawing indirect inferences from test scores, and in order to move beyond speculating about how and why individual differences might have been related to performance, it will be vital to consult the concurrent measures of the participants’ attentional processes and understanding of feedback. Moreover, Carroll (2001) refers to her prediction of restrictions on the potential effectiveness of feedback as the awareness constraint on negative evidence. The results of the present experiment may suggest that her prediction is wrong, but they do not yet indicate the reason(s) for this—whether it might be because learners can be aware of structural characteristics even if they are not represented phonologically, for example, or because awareness is not required for feedback to function as negative evidence, or for some other reason.

Carroll (2001) has asserted that feedback “has a causal role to play in the restructuring of knowledge only if the contents of the learner’s conceptual representations can be shown to be the essential element initiating change” (p. 202). Even if one is agnostic about the validity of that assertion itself, evidence of this sort for a causal role could be sought in learners’ introspective data. In the context of the current study, if it were found that some learners in the Trees-Feedback condition developed new metalinguistic notions (or activated existing ones in helpful ways) because of the information in the tree diagrams, that they then began to
express awareness of constraints governing interpretations of zibun, that they demonstrated improvement over the course of the tests in ways that were clearly related to the rules they verbalized, and that they improved to a greater extent than learners who did not express such awareness, the evidence of a causal role for the feedback would be strong. If participants in the other treatment conditions did not do these things, then the argument might seem even stronger, although it is certainly conceivable that there could be alternative routes to the same knowledge.

The fact that there were two learners who achieved perfect scores on the post-test in the Trees-Feedback condition, but not in either of the other conditions, lends some additional support to the idea that the tree diagrams may have enabled learners to reject antecedents that would have been allowed in English (i.e., they may have provided negative evidence). It does not prove this, however. For one thing, there were only 2 learners who scored 100%; for another, the data have not been analyzed according to item type. Therefore, another step for future research will be to examine the positive- and negative-evidence items separately and in comparison with each other. Of interest will be whether the participants behaved in accordance with linguistic theory, whether one item type might have been easier than another, whether this differed across the treatment conditions, whether it varied according to individual differences, and so on. Generative SLA researchers would likely also be interested in assessing the systematicity of individual learner grammars.

Finally, regarding the interfaces between theory, methodology, and pedagogy, it is crucial to discuss limitations in the representativeness of the sentences and the accessibility of the tree diagrams which were used in the current study. These issues are important for several
reasons, one of which is to underline that the results of this experiment should not be interpreted as saying much at all about the full potential of tree diagrams to be effective as a form of feedback. If anything, the findings may suggest that tree diagrams can be helpful for some learners, even when they illustrate complex, uncommon, and sometimes awkward-sounding sentence types, and are presented in a way that is deliberately obtuse.

In relation to the goal of testing Carroll’s predictions, several techniques were employed in order to balance particular morphological and syntactic characteristics, such as different case-markings and sentential positions, against each other. The outcome was a rather unique set of sentences, some of which may have sounded natural and may have been relatively easy to process (e.g., *Obaasan-ga onnanoko-ni zibun-ni tsuite-no hanashi-o shita*, ‘The grandmother told the girl a story about herself’), but others of which may have sounded contrived and may have been difficult to process (e.g., *Afumadiinejaado-ga Puuchin-ga Kimu-ni zibun-ni tsuite-no hon-o utta-to itta*, ‘Ahmadinejad said that Putin sold Kim Jong-il a book about himself’; *Zou-ni inu-ga sukina nezumi-ga zibun-o miteiru-no-ga mieta*, ‘The elephant was able to see the mouse that the dog liked staring at himself’). On top of this, in part also because of an interest in aptitude-treatment interactions, the introduction to tree diagrams was extremely limited. In the Trees-Feedback condition, after a very brief and basic overview of how the trees could be interpreted (see Chapter 4), the participants were given no further information beyond what they could gather themselves from the diagrams they received as feedback (and it may be recalled that one of the participants, who did not find the trees intuitively understandable, referred to them as the “weird spider diagrams”). Prose explanations of reflexive binding were never provided.
If the sentences had been simpler and more natural, if a more detailed explanation of the tree diagrams had been provided, and if the participants had been given any sense of what they should pay attention to in the diagrams through some sort of prior indication as to what the rules for zibun-binding might involve (e.g., simply by mentioning subject status as being relevant), it seems likely that the diagrams would have been considerably more effective. Hence, as far as the pedagogical implications of this experiment are concerned, it seems important not to draw conclusions about the magnitude of the L2 learning effects that might be obtained through the use of tree diagrams. Presented in more incremental and scaffolded ways, along with explanations of the linguistic phenomena being illustrated, they may have a good deal of untapped potential.

All of this being said, a caveat noted by Robinson (1997b) and an insight from Bruhn-Garavito (1995) are highly relevant here. Robinson points out that most pedagogic rules, including those in his experiment, represent simplifications as opposed to exhaustive descriptions: “[M]ost pedagogic rules abstract away from the full details of linguistic fact in the interest of capturing regularities which are assimilable for learners... leaving finer details... to later explanation, or to experiential learning from exposure to examples” (p. 84, note 1). Even this does not capture the scope of the problem, however. It is not simply that teachers or textbook writers choose not to present rules in their full complexity; the fact of the matter, as pointed out by Bruhn-Garavito, is this: “If ‘natural’ input underdetermines the complex target grammar, explicit input does so even more. Our understanding of language is still very tentative” (p. 80).
As alluded to in Chapters 3 and 4 and discussed in Appendix F3, several competing analyses of zibun exist in the literature on Japanese theoretical linguistics, and many of them accord pragmatic factors much more importance than purely structural considerations. The primary source of sentence types for this experiment, however, was generative SLA research, which has tended to focus on the syntax and morphology of reflexives. In relation to the information provided in the Trees-Feedback condition, this may actually amount to a fundamental problem. The tree diagrams were insufficient as full explanations of how zibun works because full explanations are not merely structural; they involve issues of perspective, empathy, logophoricity, deixis, and so on. If the assumed requirement that zibun's antecedents be c-commanding subjects is secondary, in a sense, to other factors, then focusing a set of treatment materials on it might have helped learners to deal accurately with the sentences they were tested on, but the question arises as to whether it might also have led to feelings of contradiction with intuitions they had already been developing independently. This idea recalls some musings by Thomas at the end of her 1995 article:

The test materials do not provide a way of distinguishing... structural or astructural hypotheses because of the coincidence of logophoric NPs with structurally legitimated NPs.... This is not unexpected: we have seen that formal constraints on zibun require that its antecedent be a subject, and subject NPs are most likely to control the discourse perspective. The kinds of test stimuli which might distinguish the effects of the formal vs. discourse constraints on zibun are rare and somewhat contrived. As such, they are not good candidates for use with L2 learners. I leave this as a project for future research (pp. 231-2).
In the present context, one question which arose in relation to the results associated with RQ1 (on the relative effectiveness of the different treatments) was how it could be that learners in the No-Feedback condition seemed to demonstrate some evidence of pre-post improvement, whereas learners in the Right/Wrong-Feedback condition did not. In relation to RQ3 (on aptitude-treatment interactions), one question was how it could be that grammatical sensitivity was related to post-test performance in the No-Feedback condition, but not in the Right/Wrong-Feedback condition. Another was why learners’ course levels and years of study would be positively related to post-test performance in the Right/Wrong condition, but not in the Trees-Feedback condition. (In fact, there were statistically significant negative correlations between years of study and pre-post/pre-delay gains in the Trees condition.)

The possibility was suggested that participants in the No-Feedback condition with higher grammatical sensitivity may have come up with partially correct structure-oriented hypotheses which were never contradicted by feedback since they never received any. In contrast, in the Right/Wrong condition, the feedback might have dissuaded learners with higher grammatical sensitivity from continuing to entertain partially correct structure-oriented hypotheses. As such, hypotheses of that nature might not have helped them to score higher on the post-test. In view of the additional ideas discussed in this section, one might further conjecture as follows: In the Right/Wrong condition, for learners with more Japanese experience, receiving feedback that was not explicitly structure-oriented may have allowed them to continue to entertain helpful pragmatics-oriented hypotheses, and they were able to achieve higher test scores because of a correlation between discourse perspective and subjecthood.
The ironic aspect of all of this, for a study seeking to investigate the potential benefits of providing tree diagrams, is the chance that, given enough sentences (with a wide enough variety to be representative of issues related to empathy, logophoricity, etc.), participants not receiving explicitly structure-oriented feedback might ultimately end up with more nativelike intuitions. Why? Not because learners are unable to make use of trees as feedback, but because the researchers themselves have got the emphasis wrong.

Unfortunately, it does seem possible that giving learners inaccurate pedagogical rules might sometimes have long-lasting detrimental effects. Based on an empirical study comparing highly advanced instructed and naturalistic L2 learners against native speakers, Rothman (2008) has suggested that “pedagogical simplifications... form a separate system of learned knowledge [which] can override linguistic competence (the generative system) of the L2 learner at the level of performance” (p. 85), “perhaps indefinitely” (p. 101). Rothman remains an advocate of explicit instruction, and he states that there can be valid reasons for not presenting L2 learners with theoretically sophisticated formal-linguistic explanations. In his opinion, “[i]t goes without saying that teaching adult learners of a foreign language in an explicit manner is beneficial for their success” (p. 101), and rules must often be simplified “in order to make them manageable and accessible” (p. 85). However, noting that “not even one naturalistic learner demonstrated the observed pattern of target-deviancy that the vast majority of tutored learners did” in his study (p. 99), he emphasizes that teachers should attempt to present rules that truly describe underlying linguistic competence, “to the extent that this is possible” (p. 100). Particularly in light of these findings and recommendations, it is hoped that the present experiment did not provide the participants with inaccurate rules.
Conclusion

Experimental research into aptitude-treatment interactions in the field of SLA has begun uncovering a variety of relationships among learner characteristics and L2 performance in different types of task and instructional conditions, on different types of linguistic targets, and using different types of outcome measures (e.g., de Graaff, 1997; DeKeyser, 1993; Erlam, 2005; Goo, 2010; Lado, 2008; Révész, to appear (b); Robinson, 1997b, 2002; Sheen, 2008; Wesche, 1981). It has even been possible, through qualitative analyses, to begin drawing connections between learners’ aptitudes and specific L2 interactional processes and learning priorities (e.g., Fujii, 2005). In response to calls for more theory-based conceptions and measurements of aptitudes for L2 learning (e.g., Skehan, 2002, 2011), SLA researchers have begun to draw insights from other fields, such as cognitive and educational psychology, in order to adapt or develop more sophisticated, fine-grained, and valid approaches to assessing variables such as working memory capacity (e.g., Mackey et al., 2002; Goo, 2010); attention control (Trofimovich et al., 2007); and the ability to deal with novelty and ambiguity (Grigorenko et al., 2000), to name a few. At the same time, the use of traditional instruments for measuring language aptitude, such as the MLAT, has continued to produce insights regarding the types of abilities related to L2 performance under particular conditions for particular linguistic targets.

The more individual-difference variables, treatment conditions, and outcome measures are included in a study’s design, the greater possibility there may be for a researcher to risk not seeing the forest for the trees. However, it could be argued that this poses perhaps less intractable problems than the opposite situation of not having considered enough relevant variables. If mediating factors are not anticipated and measured, the results of a study might
appear to be counterintuitive or counter-theoretical when, in actuality, the analyses are simply not fine-grained enough and higher-order interactions should have been considered (DeKeyser, 1993). Worse yet, the results might not seem interpretable at all, or might be interpreted in the wrong way. Lado’s (2008) search for aptitude-treatment interactions involving more versus less explicit treatment conditions and a variety of cognitive abilities (aptitude, WM, L1 PSTM, L2 PSTM) indicates that interactions are not always discovered when looked for. Nonetheless, it is crucial for researchers to check.

In ATI research, there may be a danger, as Cronbach (1975) warned, of entering a ‘hall of mirrors’ that seems to extend indefinitely; in fact, Cronbach’s frustrations with attempting to identify ATIs and match learners to appropriate treatment conditions led him to advise that instead of trying to capitalize on the proactive matching of learners and treatments, one should monitor students’ responses to instruction and adjust as necessary. In the field of SLA in the twenty-first century, however, considering the theoretical and methodological advances that have been (and will be) made, and the corresponding degrees of focus we should be able to achieve in our choices of linguistic targets and instructional conditions, it seems reasonable to believe that, in addition to making reactive adjustments, we will eventually be able to be proactive in ways that Cronbach found elusive. One simple and intuitive outcome of the present study, for instance, is a preliminary corroboration of the hypothesis that learners with stronger visual short-term memory might be better able to make use of tree diagrams as feedback on linguistic phenomena that involve underlying hierarchical structure. Knowing this, researchers could design further studies to explore how visual representations can be used as effectively as possible.
At the very least, what can be recommended is that SLA researchers should examine narrowly defined relationships before drawing general conclusions. This will be valuable for a variety of reasons— theoretical, methodological, and pedagogical. In terms of SLA theory, discovering which individual differences are related to the learning of different types of linguistic targets under different circumstances can help us to understand more about the processes of SLA. When ATIs are found, they provide more insight into learners’ approaches to task conditions, the abilities they make use of, and the nature of the learning targets from their points of view, than researchers can gain from analyzing any of those separately (DeKeyser, 2003). Thanks to researchers such as Skehan (2002) and Robinson (2005a), to name only two, there are several theoretically rigorous and well justified hypotheses in this area. Skehan (2002) has proposed a variety of abilities (e.g., auditory segmentation, attention management, inductive learning ability, retrieval processes) which may be related to specific stages of SLA processing (e.g., noticing, complexifying, becoming accurate, creating a repertoire, automatizing), and in Robinson’s (2005a) proposals about aptitude complexes, he has taken this further to hypothesize links between combinations of abilities (e.g., memory for contingent speech, metalinguistic rule rehearsal), SLA constructs (e.g., remembering the forms of recasts, making cognitive comparisons), and task-design features (e.g., planning time, reasoning demands). Theory-based reasoning may not always pan out as expected, even for the originators of the theories; however, the ATIs that do emerge can help us to refine our theories and methodologies.

As discussed in Chapter 2, some of the aptitude-treatment interactions involving grammatical sensitivity in Robinson’s (1997b) experiment appeared to indicate that his ‘implicit
learning’ condition had not been approached so implicitly after all. In Robinson’s (2002) study, some conclusions about incidental learning that had seemed to be supported initially turned out to be somewhat inaccurate once the results for the different learning targets were examined separately. In the case of de Graaff’s (1997) research, what the researcher originally thought would be a ‘simple’ morphological target appeared ultimately not to be, and the way he realized this was by noticing that the learners in the explicit versus implicit conditions performed much more differently than he had expected for that target. In the present study, unexpected relationships and lacks of relationships (e.g., involving grammatical sensitivity, years of study, and course levels in the Right/Wrong- versus Trees-Feedback conditions)—to the extent they have been interpreted correctly—have seemed to shed indirect light on the different types of approaches learners might have been taking. Analyses of learners’ introspective data will likely provide verifications, refutations, and new insights as well.

Methodologically speaking, reflecting on these studies produces the realization that ATIs can help SLA researchers to evaluate how accurate our labels of experimental conditions and linguistic targets are. In doing so, we should be able to improve the operationalizations of important theoretical constructs in future research. Just as UG theorists have seemed sometimes to let the idea of morphological ‘triggers’ influence their thinking even when they would not knowingly commit themselves to the assumptions they entailed (see Chapter 3), researchers interested in implicit and explicit learning have seemed sometimes to draw conclusions about theoretical constructs that have not successfully been actualized in a study. Robinson (1997b), for instance, might have assisted his readers (and his own thought processes) by ceasing to refer to an ‘implicit learning’ condition. De Graaff (1997) might have assisted his
readers by putting scare quotes around the word ‘simple’ when discussing his morphological
target of noun pluralization. The SLA research community should not be drawing conclusions
about implicit learning or simple targets unless our experimental conditions and chosen
linguistic targets actually merit those labels. Aptitude-treatment interactions can help us to
decide whether or not they do.

Where L2 pedagogy is concerned, what we find out from all of this research, both
incrementally and over the longer term, may indeed make us better equipped to tailor
instructional conditions to learners’ characteristics. Many SLA researchers are interested in
these issues because of their potential to drive theory forward, but L2 practitioners are
interested in them because they have students in their classrooms who might benefit from the
insights generated. These groups and their interests are not mutually exclusive, of course.

The next steps for this research will involve performing qualitative analyses of the
concurrent verbalizations produced by the participants during the treatment activities,
attempting to link specific features of those verbalizations both to learning processes and to
individual differences such as grammatical sensitivity and metalinguistic knowledge, and tracing
each of these relationships through to the learners’ performance with specific aspects of the
linguistic target. Analyses of aptitude-treatment interactions can allow for indirect inferences to
be made about the approaches taken to learning under different conditions by learners with
different profiles of abilities, but online measures of what learners with certain abilities are
actually thinking and doing at the process level can be expected to provide even more
illuminating insights, advances, and applications.
Appendix A
Information sheet used in recruiting volunteers for the study

You are invited to consider participating in a research study about how speakers of English acquire Japanese as a foreign language. The study involves comparing different kinds of corrections and other information in order to investigate how effective they are at promoting learning.

When you are learning a foreign language, your native language can sometimes have a fairly strong (and interesting!) influence on the learning process. Some aspects of the new language may seem easy to you, while others may seem more difficult. This experiment focuses on aspects of Japanese that are not often taught in Japanese classes, and that are different in certain respects from what you might expect based on how English works. I have designed a series of individualized computer activities to help you learn about them.

In addition to exploring the effectiveness of the different activities for learning, I am also interested in the influence of individual predispositions—that is, how different people (for example, more memory-oriented or more grammar-oriented) might respond differently to different kinds of information, and how instruction can be tailored to fit people’s profiles of abilities and preferences.

There are 3 experimental conditions in this study. If you agree to participate, you will be randomly assigned to one of them. No matter which condition you are assigned to, I will be happy to meet with you at the end of the study to explain any remaining questions you might have.

Participation (over the course of 3 weeks) will involve meeting 5 times (total) for about an hour each time. All in all, you’ll complete the following activities:

- A background questionnaire (asking primarily about your previous study of Japanese and other foreign languages)
- Measures of language aptitude, working memory, and visuo-spatial memory (the results of which will gladly be shared with you if you are interested!)
- Three learning activities, which will involve reading sentences in Japanese, looking at funny pictures of celebrities and animals, and deciding if the sentences can be used to describe the pictures
- Pre- and post-tests so that the effectiveness of the different learning activities can be assessed and compared
- Some brief questionnaires asking for your perspectives on the activities
- A chance to meet with me in case you still have any questions at the end of the experiment or in case you’d like to discuss which learning strategies might work best for you

I hope the experience will be both fun and helpful for you in your study of Japanese—not just to learn about the aspects of Japanese I am specifically targeting, but also as a chance to practice vocabulary, improve your katakana, hiragana, and kanji reading skills, and strengthen your ability to interpret advanced sentence structures.

Thank you very much for your time and help!

If you have any questions at all, please feel free to contact me (Rebecca Sachs) at redbeeka@gmail.com or 202-487-7023. I’ll be very happy to talk with you!
Appendix B
Informed consent document

GEORGETOWN UNIVERSITY
CONSENT TO PARTICIPATE IN RESEARCH INVOLVING TREATMENT

PROJECT TITLE:
Corrective feedback and the acquisition of interpretive options in a second language

PROJECT DIRECTOR
Dr. Alison Mackey
Associate Professor, Department of Linguistics
Georgetown University

PRINCIPAL INVESTIGATOR
Rebecca Sachs
TELEPHONE
(202) 487-7023

The Georgetown University Institutional Review Board (IRB) has approved this research project. For information on your rights as a research subject, call the Institutional Review Board office at 202-687-1506.

INTRODUCTION
You are invited to consider participating in a research study to investigate how native speakers of English learn to interpret sentences in Japanese as a foreign language. This form will describe the purpose and nature of the research, its possible risks and benefits, other options available to you, and your rights as a participant in the study. Please take whatever time you need to discuss the study with anyone you wish. The decision to participate, or not to participate, is yours. If you decide to participate, please be sure to sign and date the last page of this form.

WHY IS THIS RESEARCH STUDY BEING DONE?
In this research study, we are comparing different kinds of corrective feedback techniques in order to investigate how effective they are at helping English speakers to learn Japanese. We are also exploring individual differences between learners which might affect how they respond to different types of feedback. When you learn a foreign language, your native language can sometimes have an influence on the process. Some aspects of the new language will seem easy to you, while others may seem more difficult. This study focuses on aspects of Japanese that are not often taught in Japanese classes, but that may be different from what you would expect based on how English works. We have designed a series of computer activities to help you understand some of the similarities and differences between the two languages.

HOW MANY PEOPLE WILL TAKE PART IN THE STUDY?
About 70 people will take part in this study. Many will be students recruited from Japanese classes at Georgetown University; others will be students from other universities in the DC area. Participants in the study are referred to as “subjects.”
WHAT IS INVOLVED IN THE STUDY?

This study mainly involves engaging in computer activities designed to help you learn aspects of Japanese grammar. It also involves some tests of what you have learned and some measurements of various abilities that have been found to be associated with language learning.

There are three possible experimental conditions in this study. If you agree to participate, you will be randomly assigned to one of them. A computer will determine your group assignment through a process that is much like picking names out of a hat. This process is called randomization. Your chance of being in any group is one in three. If you wish to know your group assignment, Rebecca Sachs (the principal investigator) will be able to share this information with you upon completion of the study.

Participation will involve meeting with the researcher 5 times over the period of a few weeks. Each meeting will take approximately one hour (depending in part on how long you would like to spend) and, all in all, will involve the following activities, mostly computer-based:

- A background questionnaire (asking primarily about your previous study of Japanese)
- Measures of language aptitude and working memory capacity (the results of which will gladly be shared with you, if you so choose)
- Three learning activities (Half of the participants will be asked to think out loud during these learning activities. If you are one of these participants, you will be asked whether your voice can be recorded.)
- A pre-test, immediate post-test, and delayed post-test so that the effectiveness of the different learning activities can be compared
- A brief questionnaire each day so that you can share your perspectives on the computer activities you have engaged in
- A chance to meet with the principal investigator in case you still have any questions at the end of the study

HOW LONG WILL I BE IN THE STUDY?

The amount of time you spend will depend in part on you, but should take no more than about 5-8 hours of your time, spread out over a few weeks in 5 sessions lasting usually about an hour.

The investigators or sponsors may stop the study or take you out of the study at any time they judge it is in your best interest (e.g., if you experience an injury or become ill, if you need additional or different interventions, or if you do not comply with the study plan). They may also consider removing you from the study for various other reasons (e.g., if they find that you already know the language points being targeted, in which case you have no need for the activities). They can do this without your consent.

You can stop participating at any time, without penalty. This research has nothing to do with your Japanese courses, and your decision as to whether or not to participate will not have any effect on your grades or on your relationship with your Japanese instructor. We would encourage you, however, to talk with the researcher before you decide to stop participating in the study.
WHAT ARE THE RISKS OF THE STUDY?

This study involves no foreseeable risks besides whatever anxiety you might feel regarding using computers, taking routine educational tests, and receiving feedback on your use of Japanese. Attempts have been made to ensure that the activities will be enjoyable, encouraging, and helpful to you, and your responses will be kept confidential, as described below. There may, of course, be risks that we cannot predict. You should feel free to discuss any concerns you might have with the researcher and with anyone else that you wish to.

ARE THERE BENEFITS TO TAKING PART IN THE STUDY?

It is reasonable to expect that this research study will assist you in your learning of Japanese, and in particular will help you with aspects of Japanese that are not often explicitly focused on in language classes despite posing special problems for speakers of English. Of course, we cannot completely guarantee that you will personally experience benefits from participating in this study. Others may benefit in the future from the information we obtain.

WHO CAN PARTICIPATE IN THE STUDY?

This study is designed for intermediate- and advanced-level university learners of Japanese as a foreign language. Your suitability for this study will be determined by your enrollment in Japanese language courses.

WHO CANNOT PARTICIPATE IN THE STUDY?

People who have never studied Japanese or who do not speak English at a native or near-native level of proficiency cannot participate in this study.

WHAT OTHER OPTIONS ARE THERE?

If you do not wish to participate in this study, please feel free to consult with the researcher or with your course instructor regarding other activities which might help you to practice Japanese.

WHAT ABOUT CONFIDENTIALITY?

Your name will not be used when data from this study are published. Every effort will be made to keep your research records and other personal information confidential. However, we cannot guarantee absolute confidentiality. Individuals from the Georgetown University IRB, other Georgetown University offices, Federal regulatory agencies, and the primary investigator may look at records related to the study, both to assure quality control and to analyze data. Your name and any material that could identify you will remain confidential except as may be required by law.
We will take the following steps to keep information about you confidential, and to protect it from unauthorized disclosure, tampering, or damage: At the beginning of the study, you will be assigned a numerical code (e.g., Participant #42) which will be used to label all of your data. Any personal information you provide will be associated with that code, and number-name correspondences will be available, if necessary, only to the researchers/entities listed above. All data will be stored in a secure, password-protected database, and efforts will be made to ensure that identifying personal characteristics are not apparent in any of the examples which are used to illustrate trends in the data when the results of this study are published.

**WHAT ARE THE COSTS?**

Qualified study subjects will not have to pay for the study intervention.

**WILL I BE PAID FOR PARTICIPATING?**

Subjects will not be paid for participating in this study.

**WHAT IF I GET INJURED DURING MY PARTICIPATION?**

The researcher will make every effort to prevent study-related injuries and illnesses. If you are injured or become ill while you are in the study, you will receive emergency medical care. The costs of this care will be charged to you or to your health insurer. No funds have been made available by Georgetown University or its affiliates, the District of Columbia, or the Federal government to compensate you for a study-related injury or illness.

**WHAT ARE MY RIGHTS AS A RESEARCH PARTICIPANT?**

Participation in this study is entirely voluntary at all times. You have the right not to participate at all or to leave the study at any time. Deciding not to participate or choosing to leave the study will not result in any penalty or loss of benefits to which you are entitled, and it will not harm your relationship with Georgetown University or any of its employees.

If you decide to leave the study, please simply inform the researcher of your intention to withdraw. You will not be obligated to complete any further part of the study. Throughout the study, a researcher will be available to discuss your interest in remaining in the study.

**WHOM DO I CONTACT IF I HAVE QUESTIONS OR PROBLEMS?**

Call Rebecca Sachs at 202-487-7023 day or night if you have questions or problems related to the study or if you think that something unusual or unexpected is happening. Call the Georgetown University IRB Office at 202-687-1506 with any questions about your rights as a research participant.
Statement of Person Obtaining Informed Consent

I have fully explained this study to the subject. I have discussed the study’s purpose, its experimental and non-experimental procedures and interventions, the possible risks and benefits, the standard and research aspects of the study, the alternatives to participation, and the voluntary nature of participation. I have invited the subject to ask questions and have answered any questions that the subject has asked.

________________________________________  __________________________
Signature of Person Obtaining Informed Consent  Date

Consent of Subject (or Legally Authorized Representative)

I have read the information provided in this Informed Consent Document (or it was read to me by ________________). My questions were answered to my satisfaction. I voluntarily agree to participate in this study.

________________________________________  ________________________
Signature of Subject      Date

________________________________________  ________________________
Signature of Legally Authorized Representative   Date

Where Appropriate

Upon signing, the subject or the legally authorized representative will receive a copy of this form, and the original will become part of the subject’s clinical record. If there is no relevant clinical record, the original will be held in the subject’s research record.

Statement of Witness

I have personally witnessed (check as applicable):

☐ The subject (or legally authorized representative or parent or guardian) sign this informed consent document.

☐ The informed consent process involving both the subject (or legally authorized representative or parent or guardian) and the person obtaining consent.

________________________________________  Date
Witness Signature
Appendix C
Methods employed in investigating interpretations of reflexives

A variety of task types have been used to investigate L2 learners’ interpretations of reflexives, and some have been more successful than others at avoiding the problem of preferred readings. As mentioned in Chapter 4, it has been suggested that, in general, studies employing picture-identification tasks (e.g., Christie, 1992; Finer, 1991; Finer & Broselow, 1986) and multiple-choice interpretation tasks (Hirakawa, 1990; Matsumura, 1994; Thomas, 1989, 1991; Yuan, 1998) might be considered “contaminated” by the preference effect (Matsumura, 2007). However, studies employing truth-value judgment tasks (e.g., Akiyama, 2002; Hamilton, 1998; L. Jiang, 2009; Matsumura, 2007; Thomas, 1995; Wells, 1998; White, 1995b; White et al., 1996) may be less so.

Picture-identification tasks

Picture-identification tasks commonly involve presenting participants with sentences and pairs of pictures and asking them which of the two pictures illustrate(s) what each sentence can mean. This method may underestimate the types of antecedents which the participants would allow under other circumstances. In a study by Christie (1992), only 45% of native speakers’ responses indicated acceptance of theoretically allowable local non-subject antecedents in English, leading Thomas to conclude that Christie’s test procedure “[d]id not succeed in surmounting the preference for subject binding” (p. 217). In Finer’s (1991) study, only 25% of native speakers’ responses indicated ambiguous interpretations allowing both local and LD subject antecedents in Japanese; 45% bound zibun exclusively long-distance despite the fact that local subject antecedents are also possible. The current consensus is that picture-identification tasks do not elicit a valid representation of the interpretations that native speakers or L2 learners actually allow for sentences with reflexives.

Multiple-choice interpretation tasks

Many variations on multiple-choice interpretation tasks have been used in reflexives research. In Thomas’ (1989) study, for each sentence, participants were presented with 3 options. For example, for an English sentence such as Cat Stevens told Kermit a story about himself, participants would be asked whom the reflexive could refer to and given the choices of (a) Cat Stevens, (b) Kermit, and (c) Either Cat Stevens or Kermit. (Theoretically, the correct answer would be (c), either.) Using this method, Thomas found that the adult native speakers of English in her control group preferred subject antecedents in neutral monoclusal sentences (selecting subjects at a rate of 73%, objects at 2%, and either at 25%), but that many were willing to allow non-subject antecedents as well when they were pragmatically favored (20% subjects, 54% objects, 26% ambiguous).

Hirakawa (1990) used a similar type of test in a study of L1-Japanese speakers’ knowledge of English reflexives; however, considering that the reflexive zibun in Japanese can also have the speaker of the sentence as an antecedent, she provided the learners with 5 possible choices (e.g., (a) Cat Stevens, (b) Kermit, (c) Either Cat Stevens or Kermit, (d) Someone else: ______, and (e) Don’t know) and provided the native speakers with 4 possible choices (leaving out the “Don’t know” option). Importantly, the participants in Hirakawa’s experiment
were told that “the purpose of the study was to find out how [they felt]... and not to test their knowledge, so that they should relax and... [give] their initial response to each sentence” (p. 69). Hirakawa found that the Japanese native speakers overwhelmingly preferred long-distance antecedents, despite the fact that local ones are also possible. On biclausal items, 63-71% of responses were long-distance, with only about 10% indicating ambiguity. On triclausal items, it was the antecedent in the middle that was chosen most frequently (53-60%); again, the local antecedent was chosen only about 9-11% of the time, and 16-21% of responses indicated ambiguity. As in many other studies, the native speakers of English in Hirakawa’s control group showed a preference for subject antecedents (67% subject, 21% object, 12% ambiguous).

In Thomas’ (1991) experiment, the choices included (equivalents of) (a) Cat, (b) Kermit, (c) Some other person, (d) Cat or some other person, (e) Kermit or some other person, (f) Cat or Kermit, and (g) Cat or Kermit or some other person. In this study, Thomas held a special pre-task training session to try to increase participants’ awareness of ambiguous interpretations; however, she still found that only about half of the native speakers of English consistently perceived both subjects and non-subjects as possible antecedents in English (52% subject only, 0% non-subject only, 48% ambiguous), and very few of the native speakers of Japanese consistently perceived both local and LD subjects as possible antecedents in Japanese (0% local only, 60% LD only, 10% ambiguous). The rate of ambiguous interpretation in L1 English was appreciably greater than in Hirakawa’s (1990) experiment, but still less than might have been expected based on the theory that both subjects and non-subjects are allowable as antecedents for English reflexives.

Basing their approach on a method used by Lakshmanan and Teranishi (1994), MacLaughlin (1998) and Yip and Tang (1998) asked participants separately for each option whether they agreed or disagreed that it could be an antecedent. The idea was that forcing participants to reject the options they found impossible would reduce the chance that they simply failed to consider possible interpretations (Yip & Tang, 1998, p. 175). MacLaughlin’s items were formatted roughly as follows (p. 210):

(1) Tom thinks that John hates himself.
   himself can be John       Agree _____ Disagree _____
   himself can be Tom        Agree _____ Disagree _____

She did not find the approach to be particularly successful: “[W]hereas a few English controls consistently reported multiple interpretations [in cases where multiple interpretations were possible]... the majority recognized only one of two possible grammatical antecedents” (pp. 222-3, note 21). Yip and Tang’s tests, which were given in both (L2) English and (L1) Cantonese, used the following format (p. 177):

(2) John thought that Bill praised himself.
   Can ‘himself’ refer to John?   Yes No
   Can ‘himself’ refer to Bill?   Yes No
   Can ‘himself’ refer to somebody else? Yes No

Regarding the participants’ binding of monomorphemic zigei in their L1, between 48-69% of their responses indicated ambiguous interpretations; 1-9% were only local, and 11-17% were only long-distance.
In yet another variation on a multiple-choice interpretation format, Wakabayashi (1996) attempted to induce his participants to consider the possibility of each antecedent separately by listing several options (Cat Stevens, Kermit, Someone not mentioned, I don’t know) and requiring the participants to indicate not only whether or not each could be an antecedent, but also which they preferred, by numbering them in order. (In fact, it was a bit more complicated than this. If participants felt that some of the options were equally acceptable, they were instructed to assign them the same number to indicate a tie.) Wakabayashi notes that the instructions emphasized the possibility of indicating that all options were acceptable as antecedents. He nonetheless found, based on group analyses, that 54% of the native speakers’ responses indicated that only subject antecedents were allowed in English (p. 280). Analyzing the results by individual, it turned out that the responses of 67% of native speakers were either inconsistent (sometimes allowing non-subject antecedents) or consistent in allowing both subject and non-subject antecedents.

**Truth-value judgment tasks**

Truth-value judgment tasks (TVJTs) sometimes employ written stories to set a context for interpretation, sometimes pictures, and sometimes both stories and pictures. As a whole, this genre of task has been found to be more effective than picture-identification and multiple-choice tasks, but the results can differ according to the contextualization technique employed.

**Story-based**

In an experiment by White (1995b), the participants were asked to read a series of written stories, each followed by a comment, and had to indicate whether the comment was true or false in the context of the given story. The context in each story was designed to support only one interpretation of each sentence. For the target items, each of these comments was an English sentence containing a reflexive, and each was grammatical under at least one interpretation; the question was whether the learners’ grammars allowed them to have the intended interpretation. White made certain to balance the number of true and false items, as well as the number of male and female antecedents, and, as in most experiments, restricted each sentence to including possible referents of only one gender. For example:

(3) *Finite biclausal item, targeting a long-distance antecedent (correct answer: false)*

Johnny and a little boy were playing with matches. Johnny lit a match and then dropped it on the little boy’s leg. The little boy went screaming to his father and told him what had happened.

The little boy said Johnny burned himself. True or False

White’s approach to the problem of preferences was to pilot the materials beforehand and include only those items on which native speakers performed as expected. (She notes that during piloting, the native speakers quite often rejected object antecedents even when they should theoretically have been allowed and even when the context strongly favored them.)

Akiyama (2002) also used a story-based TVJT, but a difference is that the stories were written in the participants’ L1 (Japanese) on one side of a page in a test booklet, while the comments to be judged were written in their L2 (English) on the other side. Akiyama’s rationale
for designing the task in this way had two goals: (1) to ensure that participants with lower L2 proficiency would understand the stories completely, and (2) to ensure that the participants would process the meanings of the stories before they were asked for their interpretations of the sentences in English. Akiyama reasoned that printing the Japanese stories and English sentences on opposite sides of each page might prevent interference from the L1. However, the participants were allowed to flip back and read each story as many times as they felt necessary before judging the English sentence; hence, it seems reasonable to consider that they may have been in bilingual mode (Grosjean, 2001). (For the native speakers of Japanese included as a control group, everything was written in Japanese.) The instructions asked the participants to judge whether the content of each sentence could be “naturally inferred from the story” (p. 35). Analyzing the native speakers’ results, Akiyama found that both local and long-distance antecedents were accepted 68% of the time for the items with finite clauses and 92% of the time for those with nonfinite clauses (a “dramatic” improvement over the roughly 10% rate found by Hirakawa [1990]) (p. 39).

**Picture-based**

In White et al.’s (1996) study involving an experimental treatment with a pre-test/post-test design, the testing materials were in the form of a picture-based TVJT. To facilitate the participants’ understanding of the pictures, 3 characters were used: Mr. Black, Mr. Grey, and Mr. White, each of whose possessions were always supposed to be the color implied by the person’s name. For example, the sentence *White-san-ga Grey-san-ni zibun-no-pen-o watashimashita* (‘Mr. White passed self’s pen to Mr. Grey’) was presented twice, each time with a different picture. One time, the pen depicted was white, whereas the other time, it was grey. Each item was given its own full page in a test booklet.

Most of the native speakers behaved largely as predicted, except on the biclausal items targeting local antecedents. White et al. (1996) report that all 9 native speakers of Japanese consistently accepted subject antecedents in monoclausal sentences and long-distance subject antecedents in biclausal sentences. They also consistently rejected local object antecedents in monoclausal sentences, and 8 of the 9 consistently rejected LD object antecedents. However, only 4 of the 9 consistently accepted local subjects in biclausal sentences, and the group average for that item type was only 1.89 out of a possible 4 points.

**Story- and picture-based**

The TVJT employed in Thomas’ (1995) study utilized both stories and pictures for each item. Each illustrated story in Japanese was presented in two formats: using conventional Japanese orthography and also in romanized form. The two versions were presented opposite each other on facing pages of a test booklet. Each story was between 3-5 sentences long, with hand-drawn stick figures acting out the events. The statement to be judged was at the bottom of the page, and the participants were asked to circle *hai* (yes) or *iie* (no) to indicate whether the statement made sense in the context of the story.

---

45 Another reason for doing this, mentioned by L. Jiang (2009), was to “eliminate any possible attempts by the participants to somehow use the surface grammatical form of the stories as a source of aid in judging the comments” (p. 478). Jiang also used a TVJT with stories presented in the participants’ L1.
As a group, the native speakers of Japanese accepted local subjects as antecedents 93% of the time in monoclausal sentences and accepted LD subjects 89% of the time in biclausal sentences. The fact that they also unexpectedly accepted local non-subject antecedents 13% of the time in monoclausal sentences and LD non-subject antecedents 18% of the time in biclausal sentences (what Thomas calls “low but not vanishing rates”) prompted her to suggest that although there was a clear contrast between acceptance of subject and non-subject antecedents, the task may not have been a perfect reflection of the speakers’ grammars (p. 224). The individual analyses, which Thomas conducted in a series, showed that 33/34 of the native speakers consistently accepted local subject antecedents in monoclausal sentences, 29 of those 33 consistently accepted LD subject antecedents, 26 of those 29 consistently rejected LD non-subject antecedents, and 25 of those 26 also consistently rejected local non-subject antecedents in monoclausal sentences. For Thomas, the most relevant outcome was that 3 native speakers of Japanese consistently allowed LD binding to both subjects and non-subjects, a pattern considered to be unsanctioned by UG. It also means that 25/34 (74%) of the native speakers behaved overall as would be predicted by theoretical accounts of what can serve as possible antecedents for zibun (i.e., local and LD subjects, but no objects).

A final study to be discussed here, by White et al. (1997), was performed with the express purpose of comparing picture-based and story-based versions of a TVJT. The participants were L1-Japanese and L1-French adults studying L2 English at a high-intermediate or low-advanced level in an intensive summer program in Canada (plus Canadian anglophones as controls). A variety of sentence types were employed, including monoclausal sentences which targeted subject and object antecedents, and both finite and nonfinite biclausal sentences which targeted both local and LD subject antecedents. Each sentence occurred twice during the task, with a different context and therefore a different targeted interpretation each time. The story-based version was similar (if not identical) to the one used by White (1995b), described above. The picture-based version was similar to the one used by White et al. (1996); each picture-sentence pairing was presented on its own page, and all of the sentences involved the same 2 characters, Mr. Brown and Mr. Green, who were depicted as wearing shirts that matched the colors of their names. The researchers were sure to vary the characters’ locations in the pictures as well as their grammatical functions within the sentences.

When White et al. (1997) analyzed the participants’ responses according to group averages on the story task, they found that the native speakers were more accurate than the L2 learners, but that the L2 groups did not differ from each other (out of 24 possible points, the totals were as follows: L1-Japanese: 16.63, L1-French: 18.86, L1-English (NS): 22.58). On the picture task, there were no statistical differences among the groups (totals: L1-Japanese: 17.69, L1-French: 18.27, L1-English (NS): 20.43). Overall comparisons between the task types did not produce any differences.

Looking at the various sentence types separately, however, there were indications that the tasks functioned differently. For instance, whereas the results were similar across groups and tasks for the monoclausal items targeting subject antecedents (which work the same in all of the L1s), the results differed across tasks for the monoclausal items targeting object antecedents. Comparing performance with object antecedents on the picture task versus the story task, White et al. found that all 3 groups of participants tended to reject objects (wrongly) on the picture-based version and to accept them (correctly) at higher rates on the story-based
version; that is, the picture task may not have reflected the participants’ grammars with regard to a dispreferred but theoretically acceptable option. As another way of examining performance with object antecedents, White et al. looked at each group’s accuracy with subject versus object antecedents on each task type separately. They found that all groups responded more accurately with subjects than with objects on both tasks, except for the native speakers on the story-based version, suggesting (1) that the story task may have been more valid, and (2) that the L2 learners may not yet have acquired the possibility of object antecedents.

Regarding the patterns of responses on the nonfinite biclausal items targeting local subject antecedents, White et al. found that the L1-Japanese learners of English were more likely than the NSs to reject local subject antecedents (wrongly) on both tasks (as might have been expected, based on the patterns shown by Japanese native speakers in previous studies). On the finite biclausal items, the L1-Japanese learners were again more likely to reject local subjects on the story task; however, the picture task did not produce any statistically significant differences between the groups. For both finite and nonfinite items targeting LD subject antecedents, both L2 groups performed less accurately than the NSs on the story task.

Analyzing the data at the individual level, White et al. similarly found clear differences between the story-based and picture-based versions of the TVJT. On the story task, as compared to the picture task, “far more participants” in each group consistently (correctly) accepted object antecedents (story task: 58-89%, picture task: 0-21%), and far fewer consistently (incorrectly) rejected objects (story task: 9-21%, picture task: 43-69%) (p. 159). Compared to studies employing picture-identification or multiple-choice formats, in which the acceptance response rate for object antecedents as well as the percentage of participants consistently accepting object antecedents in monoclausal sentences has tended to hover around 20-25% (e.g., Finer, 1991; Hirakawa, 1990; Thomas, 1991), the story-based TVJT therefore seems to be a considerable improvement, whereas the picture-based TVJT does not.

Elaborating on the significance of their results, White et al. point out that researchers might reach completely opposite conclusions through using different task types. In their words, “Possibly, neither task provides an accurate assessment. However... the story task shows acceptance of object antecedents by both NSs and ESL learners, suggesting that, for this property at least, the results reflect linguistic competence” (p. 161). In contrast, the picture task “failed to help NSs and ESL learners override a preference for subject antecedents” (p. 161). According to the authors, this may be because pictures alone do not provide a rich enough discourse context. It may also be related to participants’ approach to the task. White et al. (1997) report that some participants mentioned having tended to read the sentences first and only then look at the picture, in which case their interpretation preferences may have been set before the picture was able to bring to mind the intended reading. In light of their findings, White et al. suggest that researchers should be careful to ensure that participants actually do process the context before judging the targeted sentence. They also advise that researchers should “seek convergent evidence from a variety of tasks in order to fully understand the nature of the interlanguage grammar” (p. 163).
Appendix D1
Example of a test answer sheet

<table>
<thead>
<tr>
<th>Participant ID#</th>
<th>Version B2</th>
<th>PRE</th>
<th>POST</th>
<th>DELAY</th>
<th>Start time:</th>
<th>End time:</th>
</tr>
</thead>
</table>

21.  はい いいえ  
この絵の状況で上のセンテンスが使えますか？

<table>
<thead>
<tr>
<th>21.</th>
<th>魔女のお姉さんより？</th>
<th>How confident do you feel?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
<td>50% 51-60% 61-70% 71-80% 81-90% 91-99% 100%</td>
</tr>
<tr>
<td></td>
<td>What was your judgment based on?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>guess intuition memory rule</td>
<td></td>
</tr>
</tbody>
</table>

22.  はい いいえ  

<table>
<thead>
<tr>
<th>22.</th>
<th>ハンクをたたいた？</th>
<th>How confident do you feel?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
<td>50% 51-60% 61-70% 71-80% 81-90% 91-99% 100%</td>
</tr>
<tr>
<td></td>
<td>What was your judgment based on?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>guess intuition memory rule</td>
<td></td>
</tr>
</tbody>
</table>

23.  はい いいえ  

<table>
<thead>
<tr>
<th>23.</th>
<th>ニックをきらわない？</th>
<th>How confident do you feel?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
<td>50% 51-60% 61-70% 71-80% 81-90% 91-99% 100%</td>
</tr>
<tr>
<td></td>
<td>What was your judgment based on?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>guess intuition memory rule</td>
<td></td>
</tr>
</tbody>
</table>

24.  はい いいえ  

<table>
<thead>
<tr>
<th>24.</th>
<th>スマーフの家の中に？</th>
<th>How confident do you feel?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
<td>50% 51-60% 61-70% 71-80% 81-90% 91-99% 100%</td>
</tr>
<tr>
<td></td>
<td>What was your judgment based on?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>guess intuition memory rule</td>
<td></td>
</tr>
</tbody>
</table>

25.  はい いいえ  

<table>
<thead>
<tr>
<th>25.</th>
<th>おじいさんの家族？</th>
<th>How confident do you feel?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
<td>50% 51-60% 61-70% 71-80% 81-90% 91-99% 100%</td>
</tr>
<tr>
<td></td>
<td>What was your judgment based on?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>guess intuition memory rule</td>
<td></td>
</tr>
</tbody>
</table>
Appendix D2
Example of a treatment answer sheet

<table>
<thead>
<tr>
<th>Participant ID#</th>
<th>Activity #2</th>
<th>TR</th>
<th>YN</th>
<th>XT</th>
<th>TA</th>
<th>Start:</th>
<th>End:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>プリトニーについてのゴシップ?</td>
<td>はい</td>
<td>いいえ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>さるをする?</td>
<td>はい</td>
<td>いいえ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>サラのにきび?</td>
<td>はい</td>
<td>いいえ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>ガールフレンドのこと?</td>
<td>はい</td>
<td>いいえ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td>ドナルドのおくんさん?</td>
<td>はい</td>
<td>いいえ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16.</td>
<td>けいさつさんの写真?</td>
<td>はい</td>
<td>いいえ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17.</td>
<td>ヒラリーについての本?</td>
<td>はい</td>
<td>いいえ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td>政治家のオバマについての本?</td>
<td>はい</td>
<td>いいえ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19.</td>
<td>ハリーのオームと?</td>
<td>はい</td>
<td>いいえ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20.</td>
<td>お母さんをきらっていた?</td>
<td>はい</td>
<td>いいえ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please circle yes or no.
Appendix E1
Instructions for the tests of zibun interpretation

Thank you very much for your participation!

This activity involves making judgments about Japanese sentences.
At the top of each slide, you’ll see a short contextualizing sentence. Then a picture will appear, along with another sentence below it.
Your job will be to decide whether the sentence at the bottom of the screen can be used to express the meaning implied by the picture and context given at the top of the screen.

There will be 30 sentences for you to judge today. All of the sentences will be possible, grammatical sentences in Japanese. The issue for you to decide will be whether each sentence at the bottom of the screen can be used with the particular meaning given.

Example: チャックは犬におこっていた。犬はチャックに悪いことをしたと思った。

チャックがけった犬が自分をきらった。
(By the way, this word means ‘situation’. Does the underlined sentence work in the situation described above?)

Then circle はい or いいえ on your answer sheet – don’t forget! 😊
Confidence judgments

After each judgment that you make, you will be asked how confident you felt about your answer. Please respond on a 7-point scale:

- ‘1’ means you have made a **guess** with no confidence at all (might as well have been a coin toss)
- ‘7’ means that you’re **absolutely certain** that your judgment is correct.

You can think of the intermediate levels of confidence as corresponding to percentages (e.g., 70% certain, 85% certain, etc.).

(You don’t have to click anything for this part; just circle a number.)

**Basis for judgments**

Please also indicate what you think formed the main basis of your judgment.

- Circle **Guess** if you think you have a 50/50 chance of being right (that is, if you might as well have flipped a coin).
- Circle **Intuition** if you believe your judgment was based on a gut feeling of intuition (that is, you have some confidence in your decision but don’t know why it’s right).
- Circle **Memory** if you believe you relied on memory for previous sentences (that is, if you recall the sentence or a portion of it from the training activities).
- Circle **Rule** if you believe you applied a rule in judging the sentence (that is, if you acquired a rule during the training activities and could verbalize it).
How to respond

Please note: These PowerPoints are (unfortunately ;) not automatically connected to a database. In order to analyze your responses, I'll need a written record of them. So, not to beat a dead horse, but please...

***Don’t forget to mark each answer on your answer sheet.***

You’ll additionally want to get in the habit of clicking the Yes/No buttons on the screen because that’s how the computer will know what sort of information to give you next week, but (especially for today) it’s important to write down your answers.

Please also be sure to write your participant number at the top of the sheet, and please record the time you start and the time you finish.

Finally, as you go through the items in order, please do not go back to change any of your answers. This is very important to ensure the validity of the research.

Thanks so much again for your participation! Have fun!!!

START
Appendix E2
Instructions for the treatments

No-Feedback condition: Slide 1

Instructions

Thanks so much (again!) for being here!!!

This activity will be similar to the one you have already done. You will be asked to decide whether or not each sentence you are shown can be used with a particular meaning to describe a picture in a given context.

Keep in mind: All of the sentences will be grammatical Japanese sentences that mean something (or, in some cases, more than one thing). The question you have to answer in each case is whether or not the sentence can be used with the specific meaning shown.

There will be 40 sentences today. It may take some time for things to start feeling like they’re falling into place, so please continue paying attention to the meanings of the sentences and try to have fun! 😊

Next

No-Feedback condition: Slide 2

A few final instructions

(😊 These are important 😊)

Before you click on the Yes and No response buttons on the screen, please be sure to mark each answer on your answer sheet.

Please go through the items in order and do not go back or change any of your answers. (This is very important to ensure the validity of the research – thanks!)

Enjoy!

Next
Thanks so much (again!) for being here!!!

This activity will be similar to the one you have already done. You will be asked to decide whether or not each sentence you are shown can be used with a particular meaning to describe a picture in a given context.

Keep in mind: All of the sentences will be grammatical Japanese sentences that mean something (or, in some cases, more than one thing). The question you have to answer in each case is whether or not the sentence can be used with the specific meaning shown.

There will be 40 sentences today. The difference is that this time the computer will tell you whether your judgments are right or wrong according to the way Japanese works. Please continue paying attention to the meanings of the sentences and try to have fun! 😊

A few final instructions…
(😊 These are important 😊)

Please DO NOT use the keyboard to advance through the slides. (You need to click the Yes and No buttons with the mouse in order to get feedback that’s appropriate to how you answered.)

Also: BEFORE YOU CLICK the Yes or No button on the screen, please be sure to mark each answer on your answer sheet. (Don’t forget!)

Please go through the items in order and do not go back or change any of your answers. (This is very important to ensure the validity of the research – thanks!)

Enjoy!
Instructions

Thanks so much (again!) for being here!!!

This activity will be similar to the one you have already done. You will be asked to decide whether or not each sentence you are shown can be used with a particular meaning to describe a picture in a given context.

Keep in mind: All of the sentences will be grammatical Japanese sentences that mean something (or, in some cases, more than one thing). The question you have to answer in each case is whether or not the sentence can be used with the specific meaning shown on the slide.

Next

Instructions (cont.)

There will be 40 sentences today. This time, the computer will tell you whether your judgments are right or wrong according to the way Japanese works. Please continue paying attention to the meanings of the sentences and try to have fun! 😊

In addition, you will be shown some information which may help you to understand why each sentence works the way it does.

Since the format of that information might be a bit unfamiliar to you, let's go through a brief explanation first…

Next

(This was followed by additional instructions presented in Chapter 4 of the main text. The instructions ended with what was used as Slide 2 in the Right/Wrong-Feedback condition.)
Appendix E3
Instructions for thinking aloud

Instructions (cont.)

Today, if you don’t mind, I’m also going to ask you to talk out loud while you are doing the computer activity.

When we are thinking about something, we sometimes talk to ourselves in our own heads. Today, I’d like you to externalize that ‘inner speech’ and speak your thoughts aloud. That is, instead of simply talking to yourself in your own head, please feel free to talk to yourself using your vocal apparatus as well!

If there is anyone else in the room, I’ll set them up with headphones or earplugs so that they’re not listening to you. (That goes for me, too.)

This isn’t a test, so please don’t worry about whether you are saying something correct or incorrect, and don’t even worry about speaking in complete sentences. The point is to get insight into different people’s thought processes, whatever they happen to be. You can speak in whichever language is more comfortable at the time – English or Japanese, or both.

Instructions (cont.)

Again, please pretend you are alone in the room and say everything that you would normally say to yourself while you are thinking. Don’t feel as though you have to explain your thoughts to anybody else or give any reasons for what you are thinking – just speak out loud whatever goes through your mind in the same way that it would normally go through your mind.

Let’s practice first! On a sheet of scrap paper, please solve the following math problem, continuously speaking your thoughts out loud while you do so.

\[
\begin{align*}
7023 \\
\times 487
\end{align*}
\]

OK! Do you have any questions? Just remember to keep talking the whole time and say whatever goes through your head!
Appendix F1
List of sentences used on the tests of zibun interpretation

Abbreviations and notes:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nom</td>
<td>nominative</td>
</tr>
<tr>
<td>Dat</td>
<td>dative</td>
</tr>
<tr>
<td>Acc</td>
<td>accusative</td>
</tr>
<tr>
<td>Loc</td>
<td>local</td>
</tr>
<tr>
<td>LD</td>
<td>long-distance</td>
</tr>
<tr>
<td>NC</td>
<td>non-c-commanding</td>
</tr>
<tr>
<td>Subj</td>
<td>subject</td>
</tr>
<tr>
<td>Obj</td>
<td>object</td>
</tr>
<tr>
<td>Same</td>
<td>the answer would be the same in English and Japanese</td>
</tr>
<tr>
<td>Y</td>
<td>the correct answer is ‘yes’</td>
</tr>
<tr>
<td>N</td>
<td>the correct answer is ‘no’</td>
</tr>
<tr>
<td>A strikethrough</td>
<td>indicates a disallowed antecedent for zibun.</td>
</tr>
<tr>
<td>Bold underlining</td>
<td>indicates that a test item targeted that noun phrase, asking whether it could serve as an antecedent for zibun.</td>
</tr>
<tr>
<td>Asterisks</td>
<td>indicate sentence types that were not practiced during the treatments (i.e., generalization items in the sense that unencountered case-markings or different types of clauses were used).</td>
</tr>
</tbody>
</table>

Some of the sentences are listed with more than one targeted antecedent. This is not because the same sentence was employed more than once on the same test, but rather because the same sentence was sometimes employed across more than one test. Whenever the same sentence was used across more than one test, a different context and picture was used, and a different antecedent was targeted.

For example, considering the monoclusal [GA NI] sentence type, the sentence Obaasan-ga onnanoko-ni zibun-nituite-no hanasi-o sita (‘The grandmother told the girl a story about herself’) was employed on Version A/A2 to test a local nominative subject antecedent, while the sentence Kanzya-ga isya-ni zibun-nituite kiita (‘The patient asked the doctor about himself’) was employed on Version A/A2 to test a dative indirect object antecedent. On each of the other test versions, one of the same sentences was used for a different purpose: Obaasan-ga onnanoko-ni zibun-nituite-no hanasi-o sita (‘The grandmother told the girl a story about herself’) was employed on Version B/B2 to test a dative indirect object antecedent, and Kanzya-ga isya-ni zibun-nituite kiita (‘The patient asked the doctor about himself’) was employed on Version C/C2 to test a local nominative subject antecedent.
List of sentences used on the tests of zibun interpretation

<table>
<thead>
<tr>
<th>Sentence type [case marking]</th>
<th>Actual sentence, presented in Romanized script (with English translation)</th>
<th>Antecedent tested</th>
<th>Type of evidence</th>
<th>Test version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monoclusal, testing local subjects and/or (indirect) objects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GA GA</strong></td>
<td>Tyaaruzu-ga <strong>Uiriam-ga</strong> zibun-no okaasan-yori sukida (Charles likes William better than self’s mother)</td>
<td>Nom loc obj</td>
<td>Negative</td>
<td>A/A2</td>
</tr>
<tr>
<td></td>
<td>Asyurii-ga <strong>warui-maiyo-ga</strong> zibun-no oneesan-yori kiraida (Ashley hates the bad witch more than self’s older sister)</td>
<td>Nom loc obj</td>
<td>Negative</td>
<td>B/B2</td>
</tr>
<tr>
<td></td>
<td>Tyenii-ga <strong>Jr-Bussyu-ga</strong> zibun-no otoosan-yori sukida (Cheney likes Bush more than self’s father)</td>
<td>Nom loc obj</td>
<td>Negative</td>
<td>C/C2</td>
</tr>
<tr>
<td><strong>GA Nih</strong></td>
<td><strong>Obaasan-ga emaneke-ni</strong> zibun-nituite-no hanasi-o sita (The grandmother told the girl a story about herself)</td>
<td>Nom loc subj</td>
<td>Same (Y)</td>
<td>A/A2</td>
</tr>
<tr>
<td></td>
<td><strong>Kanzya-ga isha-ni</strong> zibun-nituite kiita (The patient asked the doctor about himself)</td>
<td>Nom loc subj</td>
<td>Same (Y)</td>
<td>C/C2</td>
</tr>
<tr>
<td></td>
<td><strong>Roian-ga Saimon-ni</strong> zibun-no tanzyoobi paatii-no koto-nituite denwa-o sita (Ryan called Simon about self’s birthday party)</td>
<td>Nom loc subj</td>
<td>Same (Y)</td>
<td>B/B2</td>
</tr>
<tr>
<td><strong>GA O</strong></td>
<td><strong>Eimii-ga Kerii-o</strong> zibun-no ie-no mae-de matta (Amy waited for Kelly outside self’s house)</td>
<td>Nom loc subj</td>
<td>Same (Y)</td>
<td>A/A2</td>
</tr>
<tr>
<td></td>
<td><strong>Kerusoo-ga Erikku-o</strong> zibun-no kanozyo kara tasuketa (Kelso rescued Eric from self’s girlfriend)</td>
<td>Acc loc obj</td>
<td>Negative</td>
<td>A/A2</td>
</tr>
<tr>
<td></td>
<td><strong>Oneesan-ga imouto-o</strong> zibun-no tomodati-ni hometa (The older sister praised the younger sister to self’s friends)</td>
<td>Nom loc subj</td>
<td>Same (Y)</td>
<td>B/B2</td>
</tr>
<tr>
<td></td>
<td><strong>Deniro-ga Skosessi-o</strong> zibun-no tanzyoobi paatii-de waratta (Deniro laughed at Scorcese at self’s birthday party)</td>
<td>Acc loc obj</td>
<td>Negative</td>
<td>B/B2</td>
</tr>
<tr>
<td></td>
<td><strong>Taany-ga obaasan-o</strong> zibun-no ie-ni turetteitta (Tanya led the old woman to self’s house)</td>
<td>Nom loc subj</td>
<td>Same (Y)</td>
<td>C/C2</td>
</tr>
<tr>
<td></td>
<td><strong>Tyenii-ga Pattiino-o</strong> zibun-no gan-de korosita (Cheney killed Pacino with self’s gun)</td>
<td>Acc loc obj</td>
<td>Negative</td>
<td>C/C2</td>
</tr>
</tbody>
</table>
### Biclausal finite, testing long-distance subject antecedents

<table>
<thead>
<tr>
<th>GA NI</th>
<th>Doree-ga Zyon-ni zibun-no CD-ga wakaranai-to omotta (Dre thought that John didn’t understand self’s music)</th>
<th>Nom LD subj</th>
<th>Positive</th>
<th>A/A2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neko-ga inu-ni zibun-no ibiki-ga kikoeru-no o sitteiru (The cat knows that the dog can hear self’s snoring)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>B/B2</td>
</tr>
<tr>
<td></td>
<td>Poora-ga Syaron-ni zibun-no huku-no yogore-ga mieru-ka kiita (Paula asked if Sharon could see the dirty spot on self’s clothes)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>C/C2</td>
</tr>
<tr>
<td>NI GA</td>
<td>Deibu-ni Kurisu-ga zibun-o waratteiru-no ga mieta (Dave was able to see Chris laughing at himself)</td>
<td>Dat LD subj</td>
<td>Positive</td>
<td>A/A2</td>
</tr>
<tr>
<td></td>
<td>Dokutaa Iiburu-ni Skotto-ga doos-ite zibun-ni okotteiru-ka kiita (Dr. Evil didn’t understand why Scott was angry at himself)</td>
<td>Dat LD subj</td>
<td>Positive</td>
<td>B/B2</td>
</tr>
<tr>
<td></td>
<td>Kurei-ni Randii-ga doosite zibun-o sinziru-ka-ga wakatta (Clay understood why Randy believes in himself)</td>
<td>Dat LD subj</td>
<td>Positive</td>
<td>C/C2</td>
</tr>
</tbody>
</table>

### Biclausal nonfinite, testing long-distance and local subject antecedents

<table>
<thead>
<tr>
<th>GA NI*</th>
<th><em>GA NI</em> Kebin-ga Nikku-ni zibun-o kirawanai yooni itta (Kevin told Nick not to hate himself)</th>
<th>Nom LD subj</th>
<th>Positive</th>
<th>A2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sutii-bun-ga Tomu-ni zibun-no hon-o moo itido yomu yooni susumeta (Stephen recommended to Tom to read self’s book again)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>C2</td>
</tr>
<tr>
<td></td>
<td>Koizumi-ga Bussyu-ni zibun-o syokaisuru yooni settokusita (Koizumi persuaded Bush to introduce himself)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>B2</td>
</tr>
</tbody>
</table>

### Causative, testing long-distance and local subject antecedents

<table>
<thead>
<tr>
<th>GA NI</th>
<th>Uta-no sensee-ga gakusee-ni zibun-no koe-o kikaseta (The singing teacher made the student listen to self’s voice)</th>
<th>Nom LD subj</th>
<th>Positive</th>
<th>A/A2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Miruhausu-ga Baato-ni zibun-no syukudai-o saset (Bart do self’s homework)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>C/C2</td>
</tr>
<tr>
<td></td>
<td>Hoooo-ga Darai Rama-ni zibun-no syuukyo-nitu-i te hanasasete (The Pope made/let the Dalai Lama talk about self’s religion)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>B/B2</td>
</tr>
<tr>
<td><em>GA O</em></td>
<td>Sensee-ga gakusee-o zibun-no tukue-ni suwaraseta (The teacher made the student sit at self’s desk)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>A/A2</td>
</tr>
<tr>
<td></td>
<td>Onisan-ga otooto-o zibun-no heya-de benkyoo-sese-ta (The big brother made/let the little brother study in self’s room)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>C/C2</td>
</tr>
<tr>
<td></td>
<td>Bosu-ga asisutanto-o zibun-no ofisu-de tabesaseta (The boss made/let the assistant eat in self’s office)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>B/B2</td>
</tr>
</tbody>
</table>
### Biclausal finite, testing long-distance and local subject antecedents and local (indirect) objects

<table>
<thead>
<tr>
<th>GA</th>
<th>GA</th>
<th>NI</th>
<th>Biclausal finite</th>
<th>Subject</th>
<th>Antecedent</th>
<th>Same</th>
<th>Negative</th>
<th>A/A2</th>
<th>B/B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maririn-ga Snuupu-ga Hefunaa-ni zibun-no ie-de atta-to itta (Marilyn said that Snoop met Hefner in self’s house)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>C/C2</td>
<td>A/A2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sutibun-ga Zyon-ga Robu-ni zibun-nituite-no situmon-o sita-to itta (Stephen said that Jon asked Rob questions about himself)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>B/B2</td>
<td>C/C2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dorii-ga Rian-ga Kyariri zibun-no konsaato-de mita-to itta (Dolly said that LeAnn saw Carrie after self’s concert)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>A/A2</td>
<td>B/B2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bobu-ga Zyakku-ga Edo-o zibun-no ireba-de kanda-to itta (Bob said that Jack bit Ed with self’s fake teeth)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>C/C2</td>
<td>A/A2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tina-ga Hirarri-ga Sara-o zibun-no soozoo-no-naka-de tataita-to itta (Tina said that Hillary hit Sarah in self’s imagination)</td>
<td>Nom LD subj</td>
<td>Positive</td>
<td>B/B2</td>
<td>C/C2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Biclausal with relative clauses, testing non-c-commanding subjects

<table>
<thead>
<tr>
<th>GA</th>
<th>GA</th>
<th>NI</th>
<th>Biclausal with relative clauses, testing non-c-commanding subjects</th>
<th>Subject</th>
<th>Antecedent</th>
<th>Same</th>
<th>Negative</th>
<th>A/A2</th>
<th>B/B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donaudo-ga yamesasetta bizinesuman-ga zibun-no koto-o hanasita (The businessperson that Donald fired talked about himself)</td>
<td>Nom NC subj</td>
<td>Same (N)</td>
<td>A/A2</td>
<td>B/B2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fan-ga aisita yuumeizin-ga terebi-de zibun-no koto-o hanasita (The celebrity that the fan loved talked about himself on TV)</td>
<td>Nom NC subj</td>
<td>Same (N)</td>
<td>B/B2</td>
<td>C/C2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kodomo-ga sukina hoisyaga zibun-o hometa (The dentist that the child liked praised himself)</td>
<td>Nom NC subj</td>
<td>Same (N)</td>
<td>C/C2</td>
<td>A/A2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finippusu-ga tatakatta suiei sensyu-ni mizu-ni utsutta zibun-ga mieta (The swimmer that Phelps beat was able to see himself reflected in the water)</td>
<td>Nom NC subj</td>
<td>Same (N)</td>
<td>A/A2</td>
<td>B/B2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tiaridaa-ga kirainaa aissuketaa-ni zibun-no kooti-no koe-ga kikoeta (The ice skater that the cheerleader hated was able to hear self’s coach’s voice)</td>
<td>Nom NC subj</td>
<td>Same (N)</td>
<td>B/B2</td>
<td>C/C2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aanii-ga issyo-ni sundeita ruumumeeto-ni zibun-no ibiki-ga kikoeta (The roommate that Ernie lived with could hear self’s snoring)</td>
<td>Nom NC subj</td>
<td>Same (N)</td>
<td>C/C2</td>
<td>A/A2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ossanoko-ga ryoori-o saseeta okaasan-ga zibun-o hometa (The mother who made the girl cook praised herself)</td>
<td>Dat NC subj</td>
<td>Same (N)</td>
<td>A/A2</td>
<td>B/B2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ossan-ga kikoenakatta kodomo-ga zibun-no kazoku-to hanasita (The child whose voice the old man couldn’t hear talked with self’s family)</td>
<td>Dat NC subj</td>
<td>Same (N)</td>
<td>B/B2</td>
<td>C/C2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Biclausal with relative clauses, testing local (indirect) objects and non-c-commanding subjects

<table>
<thead>
<tr>
<th>GA GA</th>
<th>Rarii-ga Kenan-ga tataita komedian-ni zibun-no terebi-no syoo-nituite kiita</th>
<th>Dat loc obj</th>
<th>Negative</th>
<th>A/A2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Larry asked the comedian Conan hit about self’s TV show)</td>
<td></td>
<td></td>
<td>C/C2</td>
</tr>
<tr>
<td>GA GA</td>
<td>Eruton-ga Pikatyu-ga kiraina rappaa-ni zibun-nituite-no ongaku-o okutta</td>
<td>Dat loc obj</td>
<td>Negative</td>
<td>B/B2</td>
</tr>
<tr>
<td></td>
<td>(Elton sent the rapper Pikachu hated a song about himself)</td>
<td></td>
<td></td>
<td>A/A2</td>
</tr>
<tr>
<td>GA GA</td>
<td>Inu-ga neko-ga sukina tori-ni zibun-no bideo-o miseta</td>
<td>Dat loc obj</td>
<td>Negative</td>
<td>C/C2</td>
</tr>
<tr>
<td></td>
<td>(The dog showed the bird the cat liked its video)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Biclausal with relative clauses, testing local (indirect) objects and non-c-commanding subjects

<table>
<thead>
<tr>
<th>NI GA</th>
<th>Gaasu-ni Birii-ga kaita ongaku-ga zibun-no ie-kara kikoeta</th>
<th>Dat loc subj</th>
<th>Same (Y)</th>
<th>A/A2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Garth was able to hear the song that Billy wrote from self’s house)</td>
<td></td>
<td></td>
<td>C/C2</td>
</tr>
<tr>
<td>NI GA</td>
<td>Dori-ni Pikase-ga kaita e-ga zibun-no sutazio-no-naka-ni mieta</td>
<td>Dat loc subj</td>
<td>Same (Y)</td>
<td>B/B2</td>
</tr>
<tr>
<td></td>
<td>(Dali was able to see the picture Picasso painted inside self’s studio)</td>
<td></td>
<td></td>
<td>A/A2</td>
</tr>
<tr>
<td>NI GA</td>
<td>Ben-ni Owen-ga kiraina fan-ga zibun-no takusan-no fan-no-naka-ni mieta</td>
<td>Dat loc subj</td>
<td>Same (Y)</td>
<td>A/A2</td>
</tr>
<tr>
<td></td>
<td>(Ben was able to see the fan that Owen dislikes among self’s many fans)</td>
<td></td>
<td></td>
<td>C/C2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GA GA</th>
<th>Fabio-ga Aanii-ga tataita usagi-o zibun-no kuruma-de ie-ni okutta</th>
<th>Nom NC subj</th>
<th>Same (N)</th>
<th>A/A2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Fabio drove home the rabbit that Arnie had hit in self’s car)</td>
<td></td>
<td></td>
<td>C/C2</td>
</tr>
<tr>
<td>GA GA</td>
<td>Zyei-Zi-ga Didii-ga kiraina kasyu-o zibun-no paatii-de hometa</td>
<td>Nom NC subj</td>
<td>Same (N)</td>
<td>B/B2</td>
</tr>
<tr>
<td></td>
<td>(Jay-Z praised the singer Diddy dislikes at self’s party)</td>
<td></td>
<td></td>
<td>A/A2</td>
</tr>
<tr>
<td>GA GA</td>
<td>Bosu-ga Dowaito-ga sukina Zimu-o zibun-no ofisu-de okotta</td>
<td>Nom NC subj</td>
<td>Same (N)</td>
<td>C/C2</td>
</tr>
<tr>
<td></td>
<td>(The boss yelled at the Jim that Dwight liked in self’s office)</td>
<td></td>
<td></td>
<td>B/B2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GA GA</th>
<th>Ruizi-ni Bauzaa-ga kiraina onissan-ga zibun-no kuruma-no-naka-ni mieta</th>
<th>Nom loc subj</th>
<th>Negative</th>
<th>A/A2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Luigi was able to see the brother that Bowser hated in self’s car)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GA GA</th>
<th>Gaagameru-ni neko-ga kiraina Sumaahu-ga zibun-no ie-no-naka-ni mieta</th>
<th>Nom loc subj</th>
<th>Negative</th>
<th>B/B2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Gargamel was able to see the Smurf that the cat hated in self’s house)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GA GA</th>
<th>Raiano-ga Mamura-ga atta dorakyu-ga zibun-no ie-no-naka-ni mieta</th>
<th>Nom loc subj</th>
<th>Negative</th>
<th>C/C2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Liono was able to see the vampire that Mumra met in self’s house)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GA GA</th>
<th>Papa Sumaahu-ni akatan-o turuteitta otoo-ga zibun-no ie-no-naka-ni mieta</th>
<th>Nom loc subj</th>
<th>Negative</th>
<th>A/A2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Papa Smurf was able to see the man who took away the baby inside self’s house)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GA GA</th>
<th>Raiano-ni Mamura-ga nakasetu dorakyu-ga zibun-no ie-no-naka-ni mieta</th>
<th>Nom loc subj</th>
<th>Negative</th>
<th>B/B2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Liono was able to see the vampire that made Mumra cry in self’s house)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Mario-ni Boushee-o tateita soroote-ga zibun-no kuruma-no-naka-ni mieta
(Mario was able to see the brother that hit Bowser in self’s car)

| Triclausal finite with relative clauses, testing long-distance subjects |
|---------------------------------|----------------|----------------|
| **NI GA GA**                    | Oziisan-ni gakusee-ga tateita sensee-ga zibun-o semeta koto-ga wakaranakatta |
| (The old man didn’t understand why the teacher the student hit blamed himself) | Dat LD subj | Positive | A/A2 |
| **Samurai-ni sensee-ga aisita ninzya-ga doosite zibun-o sinzita-ka-ga wakatta** |
| (The samurai understood why the ninja that the master loved trusted himself) | Dat LD subj | Positive | B/B2 |
| **Zoo-ni inu-ga suquina nezumi-ga zibun-o miteiru-no-ga mieta** |
| (The elephant saw the mouse that the dog liked staring at himself) | Dat LD subj | Positive | C/C2 |
| **Serapisuto-ni okaasan-o suquina onnanohito-ga doosite zibun-o kiraina-no-ka-ga wakaranakatta** |
| (The therapist didn’t understand why the woman who likes her mother hates herself) | Dat LD subj | Positive | A/A2 |
| **Hanku-ni Buttheadde-o suquina tomodati-ga doosite zibun-o tataita-ka-ga wakaranakatta** |
| (Hank didn’t understand why the friend who liked Butthead hit himself) | Dat LD subj | Positive | B/B2 |
| **Anzyeriina-ni Zyenifaa-o suquina onnanohito-ga zibun-o kiraida-to wakatta** |
| (Angelina understood why the woman who liked Jennifer hated herself) | Dat LD subj | Positive | C/C2 |
## Appendix F2
List of sentences used in the treatments

<table>
<thead>
<tr>
<th>Sentence type [case marking]</th>
<th>Actual sentence, presented in Romanized script (with English translation)</th>
<th>Type of antecedent</th>
<th>Type of evidence</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monoclausal, testing local subjects and (indirect) objects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| **GA GA** | Kyameron-ga *titi-ga* zibun-no kuruma-yori hosii  
(Cameron wants a father more than self’s car) | Nom loc subj | Same (Y) | 1 |
| | Tyenii-ga *Baiden-ga* zibun-no bosu-yori kiraida  
(Cheney hates Biden more than self’s boss) | Nom loc subj | Same (Y) | 1 |
| | Maatin-ga *Wiru-ga* zibun-no kodomo-yori sukida  
(Martin likes Will more than self’s child) | Nom loc obj | Negative | 1 |
| | Emeriru-ga *aironsyeh-ga* zibun-no tabemono-yori sukida  
(Emeril likes the Iron Chef more than self’s food) | Nom loc obj | Negative | 1 |
| | O J-ga *tonodati-ga* zibun-no okane-yori hosii  
(OJ wants a friend more than self’s money) | Nom loc subj | Same (Y) | 2 |
| | Roozii-ga *Donarudo-ga* zibun-no okusan-yori kirai  
(Rosie hates Donald more than self’s wife) | Nom loc obj | Negative | 2 |
| | Buritonii-ga *Parisu-ga* zibun-no inu-yori sukida  
(Britney likes Paris more than self’s dog) | Nom loc obj | Negative | 2 |
| | Madonna-ga *Pamema-ga* zibun-no mae-no otto-yori sukida  
(Madonna likes Pamela better than self’s previous husband) | Nom loc subj | Same (Y) | 3 |
| | Makku-gai *PC-gai-ga* zibun-no kompyuta-yori sukida  
(Mac-guy likes PC-guy more than self’s computer) | Nom loc obj | Negative | 3 |
| | Konan-ga *Donarudo-ga* zibun-no kami-yori kirai  
(Conan hates Donald more than self’s hair) | Nom loc obj | Negative | 3 |
| **GA NI** | Zyenifaa-ga *Syakiira-ni* zibun-no syasin-o miseta  
(Jennifer showed Shakira self’s picture) | Nom loc subj | Same (Y) | 1 |
| | Ottosan-ga *musuko-ni* zibun-nituite-no tegami-o kaita  
(The father wrote the son a letter about himself) | Nom loc subj | Same (Y) | 1 |
| | Isya-ga *kanzya-ni* zibun-no koto-o hanasita  
(The doctor told the patient some things about himself) | Nom loc subj | Same (Y) | 1 |
<table>
<thead>
<tr>
<th>Biclausal finite, testing long-distance and local subject antecedents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GA GA</strong></td>
</tr>
<tr>
<td><em>Roozii-ga Uppi-ga zibun-no koto-hanasi-to itta</em></td>
</tr>
<tr>
<td>(Rosie said that Whoopi talked about herself)</td>
</tr>
<tr>
<td>Nom LD subj  Positive  Same (Y)  1</td>
</tr>
<tr>
<td><strong>GA NI</strong></td>
</tr>
<tr>
<td><em>Saimon-ga Gaafankuru-ni zibun-utoa-no-ga kikoenai-to omotta</em></td>
</tr>
<tr>
<td>(Simon thought that Garfunkel couldn’t hear self’s singing)</td>
</tr>
<tr>
<td>Nom LD subj  Positive  Same (Y)  1</td>
</tr>
<tr>
<td><em>Erizabesu-ga Kamira-ni zibun-onaka-no oto-ga kikoeru-to omotta</em></td>
</tr>
<tr>
<td>(Elizabeth thought Camilla was able to hear self’s stomach)</td>
</tr>
<tr>
<td>Nom LD subj  Positive  Same (Y)  1</td>
</tr>
<tr>
<td><em>Gaaruhrurendo-ga booihurendo-ni zibun-koto-ga rikaidekiru-to omowanakatta</em></td>
</tr>
<tr>
<td>(The girlfriend didn’t think that the boyfriend could understand self’s things)</td>
</tr>
<tr>
<td>Nom LD subj  Positive  Same (Y)  2</td>
</tr>
<tr>
<td><em>Obaasan-ga sirayukihime-ni zibun-ga mieta-to omotta</em></td>
</tr>
<tr>
<td>(The old woman thought that Snow White saw herself)</td>
</tr>
<tr>
<td>Nom LD subj  Positive  Same (Y)  3</td>
</tr>
<tr>
<td><strong>NI GA</strong></td>
</tr>
<tr>
<td><em>Rinzii-ni Parisu-ga zibun-no kao-o miteiru-no-ga mieta</em></td>
</tr>
<tr>
<td>(Lindsay saw Paris looking at self’s face)</td>
</tr>
<tr>
<td>Dat LD subj  Positive  Same (Y)  1</td>
</tr>
<tr>
<td><em>Otokonohito-ni doo site inu-ga zibun-o kanda-no-ka ga wakaranakatta</em></td>
</tr>
<tr>
<td>(The man didn’t understand why the dog bit himself)</td>
</tr>
<tr>
<td>Dat LD subj  Positive  Same (Y)  1</td>
</tr>
<tr>
<td><em>Zyehu-ni Zimu-ga zibun-ni hansateiru-no-ga kikoeta</em></td>
</tr>
<tr>
<td>(Jeff was able to hear Jim talking to himself)</td>
</tr>
<tr>
<td>Dat LD subj  Positive  Same (Y)  2</td>
</tr>
<tr>
<td><strong>Obaamia-ni Makein-ga zibun-no koto-o hansateiru-noga kikoeta</strong></td>
</tr>
<tr>
<td>(Obama was able to hear McCain talking about himself)</td>
</tr>
<tr>
<td>Dat LD subj  Positive  Same (Y)  3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Causatives, testing long-distance and local subject antecedents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GA NI</strong></td>
</tr>
<tr>
<td><em>Inu-ga neko-ni zibun-no tabemono-o tottekosasete</em></td>
</tr>
<tr>
<td>(The dog made the cat fetch self’s food)</td>
</tr>
<tr>
<td>Nom LD subj  Positive  Same (Y)  1</td>
</tr>
<tr>
<td>*<em>Madonna-ga Buritonii-ni zibun-no bideo-o misasete</em></td>
</tr>
<tr>
<td>(Madonna made Britney watch self’s video)</td>
</tr>
<tr>
<td>Nom LD subj  Positive  Same (Y)  1</td>
</tr>
<tr>
<td>Sentence</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Okaasan-ga kodomo-ni zibun-no mono-o kawaseta</td>
</tr>
<tr>
<td>Sensee-ga saru-ni zibun-o keraseta</td>
</tr>
<tr>
<td>Ron-ga Harii-o zibun-no oomu[sic]-to asobaseta</td>
</tr>
<tr>
<td>Pinku-ga Seriinu-ni zibun-no uta-o utawaseta</td>
</tr>
<tr>
<td>Tonii-ga Kaato-ni zibun-no kuruma-o naosasetaa</td>
</tr>
<tr>
<td>Anzyeriina-ga Keitii-ni zibun-no kodomo-dake sikaraseta</td>
</tr>
<tr>
<td>Roddikku-ga Nadaru-ga Federaa-ni zibun-no kuni-de katta-to itta</td>
</tr>
<tr>
<td>Hirarii-ga Missyreu-ga Sara-ni zibun-nituite-no hon-o watasita to itta</td>
</tr>
<tr>
<td>Rinzii-ga Parisu-ga Buritonii-ni zibun-no gosippu-o itta-to itta</td>
</tr>
<tr>
<td>Kurisutiiina-ga Mairii-ga Riano-ni zibun-no bideo-o miseta-to itta</td>
</tr>
<tr>
<td>Dan-ga Maikeru-ga Rozyaa-ni zibun-no zyoozuna supootu-de katta to itta</td>
</tr>
<tr>
<td>Oya-ga okotta kodomo-ga zibun-o semeru</td>
</tr>
<tr>
<td>Chuck-ga ketta inu-ga zibun-o kirtta</td>
</tr>
<tr>
<td><strong>Kedemo-ga</strong> tataita titi-ga zibun-o semeta</td>
</tr>
<tr>
<td>---------------------------------------------</td>
</tr>
<tr>
<td><em>(The father that the child hit blamed himself)</em></td>
</tr>
<tr>
<td><strong>Tiaariidaa-ga</strong> bakanisita sukeetaa-ga zibun-no tomodati-no moe-de naita</td>
</tr>
<tr>
<td><em>(The ice skater that the cheerleader cried in front of self’s friends)</em></td>
</tr>
<tr>
<td><strong>GA NI</strong> Kyasupaa-ga kowagaraseto otokonohito-ni kagami-ni ututta zibun-ga mieta</td>
</tr>
<tr>
<td><em>(The man that Casper frightened was able to see himself reflected in the mirror)</em></td>
</tr>
<tr>
<td><strong>Sensee-ga</strong> sukina gakusee-ni zibun-no benkyoo-no naiyo-ga wakatta</td>
</tr>
<tr>
<td><em>(The student that the teacher liked was able to understand self’s study content)</em></td>
</tr>
<tr>
<td><strong>Inu-ga</strong> sukina neko-ni zibun-no onaka-no oto ga kikoeta</td>
</tr>
<tr>
<td><em>(The cat that the dog liked could hear self’s stomach)</em></td>
</tr>
<tr>
<td><strong>NI GA</strong> Sara-ga sukina kasyu-ni zibun-no nikibi-ga sukosi mieta</td>
</tr>
<tr>
<td><em>(The singer Sarah liked saw self’s pimple a little)</em></td>
</tr>
<tr>
<td><strong>Eren-ga</strong> intabyuu shita tomodati-ni zibun-no hon-no imi-ga wakatteita</td>
</tr>
<tr>
<td><em>(The friend that Ellen interviewed understood the meaning of self’s book)</em></td>
</tr>
<tr>
<td><strong>Oosutin-ga</strong> tukamaeta warui hito-ni zibun-no tanzyoobi-ga wakaranakatta</td>
</tr>
<tr>
<td><em>(The bad guy that Austin caught didn’t know self’s birthday)</em></td>
</tr>
<tr>
<td><strong>NI GA</strong> Bussyu-ni Tyelini-ga kaita hon-no imi-ga zibun-no okusan-ni kii-temo wakaranakatta</td>
</tr>
<tr>
<td><em>(Bush wasn’t able to understand the meaning of the book that Cheney wrote, even asking self’s wife)</em></td>
</tr>
<tr>
<td><strong>Kyaku-ni sutyyuwaadesu-ga</strong> ireta bakudan-ga zibun-no suutkeesu-no-naka-ni nieta</td>
</tr>
<tr>
<td><em>(The customer was able to see the bomb that the stewardess placed in self’s suitcase)</em></td>
</tr>
<tr>
<td><strong>Opera-ni Rosii-ga</strong> itta gosippu-ga zibun-no tomodati-ni kii-temo wakaranakatta</td>
</tr>
<tr>
<td><em>(Oprah didn’t understand the gossip Rosie told, even asking self’s friend)</em></td>
</tr>
<tr>
<td><strong>Buraddo-ni Zyoozi-ga</strong> totta syasin-ga zibun-no koban-no-naka-ni nieta</td>
</tr>
<tr>
<td><em>(Brad was able to see the photo that George took inside self’s bag)</em></td>
</tr>
<tr>
<td><strong>Zyakku-ni Wiru-ga</strong> itta zyooku-ga zibun-no gaaruhurendo-ni kii-temo wakaranakatta</td>
</tr>
<tr>
<td><em>(Jack didn’t understand the joke that Will told even asking self’s girlfriend)</em></td>
</tr>
<tr>
<td><strong>NI GA</strong> Neko-ni okorareta nezumi-ga zibun-no tabemono-o otosita</td>
</tr>
<tr>
<td><em>(The mouse that was scolded by the cat dropped self’s food)</em></td>
</tr>
<tr>
<td><strong>Aburiru-ni bananisareta kasyu-ga</strong> zibun-no ie-ni hi-o tuketa</td>
</tr>
<tr>
<td><em>(The singer who was insulted by Avril set fire to self’s house)</em></td>
</tr>
<tr>
<td>Clause</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td><strong>Okaasan-ni rikaidekinakatta onnanoko-ga zibun-o kiratteita</strong> (The girl whom the mother couldn’t understand hated herself)</td>
</tr>
<tr>
<td><strong>Haiyuu-ni adobaisu-o ageta direktua-ga zibun-no eiga-o tanosinda</strong> (The director who gave the actor advice enjoyed self’s movies)</td>
</tr>
<tr>
<td><strong>Sarukozni mieta nezumi-ga zibun-no tiizu-o tabeta</strong> (The mouse that Sarkozy saw ate self’s cheese)</td>
</tr>
<tr>
<td><strong>Usagi-ni tiizu-o ageta nezumi-ga zibun-no ie-ni haitta</strong> (The mouse that gave the rabbit cheese entered self’s house)</td>
</tr>
</tbody>
</table>

Biclausal with relative clauses, testing local subjects and (indirect) objects and non-c-commanding subjects

<table>
<thead>
<tr>
<th>Clause</th>
<th>Translation</th>
<th>Subject</th>
<th>Object</th>
<th>Non-commanding Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bussyu-ga Tyenni-ga kirainasezika-ni zibun-nituite-no hon-o okutta</strong> (Bush sent the politician Cheney hated a book about himself)</td>
<td>Nom loc subj</td>
<td>Same (Y)</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><strong>Syuroodaa-ga Tyawai-ga sukina inu-ni zibun-nituite-no ongaku-o hiita</strong> (Schroeder played the dog that Charlie liked some music about himself)</td>
<td>Dat loc obj</td>
<td>Negative</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Triclausal finite with relative clauses, testing long-distance and local subjects and non-c-commanding subjects

<table>
<thead>
<tr>
<th>Clause</th>
<th>Translation</th>
<th>Subject</th>
<th>Object</th>
<th>Non-commanding Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rokkusuta-ni keesatukan-ga tometa syasinka-ga zibun-no syasin-o toru-no-ga mieta</strong> (The rockstar saw the photographer that the policeman stopped taking a picture of himself)</td>
<td>Dat LD subj</td>
<td>Positive</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><strong>Operakasyu-ni poppukasyu-ga sukinazyazzukasyu-ga zibun-no uta-o utau-no-ga kikoeta</strong> (The opera singer heard the jazz singer that the pop singer liked singing self’s song)</td>
<td>Dat LD subj</td>
<td>Positive</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><strong>Bebisittaa-ni okaasan-ga okotta kodomo-ga zibun-nituite hanasu-no ga kikoeta</strong> (The babysitter was able to hear the child that the mother yelled at talking about herself)</td>
<td>Dat LD subj</td>
<td>Positive</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>
Appendix F3
Characteristics of zibun

The Japanese language has three types of reflexives: (1) monomorphemic zibun (‘self’), (2) zibun-zisin (‘self-self’), and (3) variations on kare-zisin (‘he-self’), composed of a pronominal form and a morpheme meaning ‘self’ (Aikawa, 2002, p. 154). Zibun is considered to be the most representative reflexive in Japanese, as well as being the most commonly used (Aikawa, 2002, p. 156; Akiyama, 2002, p. 28).

In addition to the requirement of a c-command relationship between antecedents and reflexives discussed in Chapter 4, another similarity between English reflexives (such as himself) and Japanese zibun is that both can serve as bound variables, as shown in example (1), adapted from Aikawa (2002, p. 163):

\[(1) \quad \text{a. Everyone, believes good things about himself;} \]
\[\text{b. Daremoi-ga [zibuni-ga sono siken-ni gookaku-suru]-to sinzi-te iru} \]
\[\quad \text{Everyone-NOM self-NOM that exam-DAT pass-COMP believe} \]
\[\quad \text{‘Everyone, believes that self will pass that exam’} \]

The quantifiers everyone/daremo do not refer to specific individuals; thus, the reflexives cannot be co-referential with them, per se. In (1a) and (1b), the interpretation is that for every person one might consider, the predicate is true of that person. Bound readings are possible for both himself and zibun, as well as for pronouns in English (e.g., he, she, his, her), but not for all pronouns in Japanese (e.g., kare ‘he’).

Zibun and English reflexives are also similar in that neither can take split antecedents, as shown in (2), adapted from Aikawa (2002, p. 159):

\[(2) \quad \text{a. *Sakiko, told Motoko, about themselves;} \]
\[\text{b. *Sakiko-ga Motoko-ni zibun-no koto-nituite hanasita} \]
\[\quad \text{Sakiko-NOM Motoko-DAT self-GEN thing-about told} \]

Beyond these similarities, there are a variety of ways in which zibun differs from English reflexives (apart from the differences in binding distance and orientation discussed in Chapters 3 and 4). For instance, zibun can be employed to refer to antecedents of any gender, person, or number, as long as they are animate. (In Japanese, it would not be possible to use zibun to say that history repeats itself, for example, since history is inanimate: *Rekisii-wa zibun-o kurikaesu [history-TOP self-ACC repeat] [Kuno, 1972, p. 178, cited in Aikawa, 2002, p. 157].) In English, on the other hand, although reflexive words are restricted in the sense that they must agree in gender, person, and number with their antecedents (e.g., herself, myself, yourself, ourselves, etc.), there are versions that can be used with inanimate referents: itself and themselves.

Other differences involve genitive case-marking, embedded subjects, prenominal modifiers, and discourse binding. Consider the following examples adapted from Tsujimura (2007) and Aikawa (2002). In each case, the English sentence given in (a) is ungrammatical, whereas the Japanese equivalent in (b) is grammatical.
(3)  a. *Hironoi studied in herself’s room
   b. *Hirono-ga zibun-no heya-de benkyoo-sita
      Hirono-NOM self-GEN room-in studied
      ‘Hirono studied in her (self’s) room’ (Tsujimura, 2007, p. 259)
(4)  a. *Kimi believes that [herself, is famous]
   b. Kimi-ga [zibun-ga yuumee-da]-to sinziteiru
      Kimi-NOM self-NOM famous-is-COMP believe
      ‘Kimi believes that she (self) is famous’ (Tsujimura, 2007, p. 262)
(5)  a. *Ayako saw a new herself in the mirror
   b. Ayako-ga kagami-no naka-ni atarasii zibuni-o mituketa
      Ayako-NOM mirror-GEN inside-in new self-ACC found
      ‘Ayako found a new self in the mirror’ (Tsujimura, 2007, p. 263)
(6)  a. Speaker A: Did Kumiko, send someone there?
   Speaker B: *No, herself (=Kumiko) went.
   b. Speaker A: *Kumiko-ga dareka-o soko-ni okutta n-desu-ka?
      Speaker B: Iie, zibuni-ga itta n-desu. (Fukui, 1984, p. 40; Aikawa, 2002, p. 157)
(7)  a. *Shozo blamed myself.
   b. Syoozoo-ga zibun-o semeta
      Shozo-NOM self-ACC blamed
      ‘Shozo blamed himself/me’ (Aikawa, 2002, p. 158)

What these examples show is that, in contrast to English reflexives, zibun can be used as a possessive (3), appear in the subject position of an embedded clause (4), accept prenominal modifiers (5), participate in discourse binding (6), and, in some dialects, be used to refer to the speaker or the hearer (7). Table F3.1, taken from Aikawa (2002, p. 159), summarizes the similarities and differences reviewed above.

Table F3.1. Summary of similarities and differences between zibun and English reflexives

<table>
<thead>
<tr>
<th></th>
<th>Zibun</th>
<th>English reflexives</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Phi-feature specification</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>b. Animacy requirement on the antecedent</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>c. Local binding</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>d. Long-distance binding</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>e. Subject orientation</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>f. Discourse-binding</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>g. Pronominal usage of “I”/“You”</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>h. Possessor position or subject position</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>i. Modification</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>j. C-command requirement</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>k. Split antecedent</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Controversies

As explained in Chapter 4, all of the sentences which were included in the testing and treatment materials of this experiment were compatible with a structure-based analysis of zibun-binding, but it has been argued that purely syntactic constraints cannot fully account for zibun’s binding behavior in the Japanese language as a whole. According to Aikawa (2002), “the subject antecedent condition on zibun-binding faces many counterexamples and its validity is still controversial” (p. 161); several researchers (e.g., Kuno, 1973, 1978; Kuroda, 1973; Kameyama, 1984, 1985; Iida & Sells, 1988; Sells, 1987; Iida, 1996) have contended that semantic and pragmatic factors (such as empathy and perspective) must also be considered.

Aikawa (2002) reviews a number of situations in which the subjecthood requirement can be “suspended,” as she puts it, and zibun can refer to a non-subject antecedent (p. 170). Among them are sentences involving reported speech, logophoricity, deixis, and the nuances associated with particular types of verbs. So that the reader can evaluate the import of these issues, some of the more striking examples which seem relevant to the types of sentences used in this experiment will be presented below.

Kuno (1973) has argued that zibun can refer to a non-subject if the clause containing zibun can be interpreted as describing the antecedent’s internal feelings or thoughts. This sometimes occurs in reported-speech contexts. For example, consider the sentences below:

    Taro-NOM Ziro-DAT Hanako-NOM self-ACC hate COMP said ‘Taro said to Ziroi (that): ‘Hanako hates me.’

    Taro-TOP Ziro-from Hanako-NOM self-ACC hate COMP heard ‘Taro heard from Ziro (that): ‘Hanako hates me.’

Unlike in English, a complement clause headed by the word to (‘that’) in Japanese can be used to express a direct quotation or expression of internal thoughts (Aikawa, 2002, p. 188, note 23). In sentence (8), which is very similar to several of the target sentences employed in the present experiment, zibun cannot refer to the dative-marked indirect object Ziro; the reported feelings make sense only as belonging to the subject Taro, who was performing the act of speaking in the sentence. However, in sentence (9), which differs from (8) only in that Ziroo-ni itta (‘said to Ziro’) is replaced with Ziroo-kara kiita (‘heard from Ziro’), the reflexive zibun can refer to the postpositional object Ziro, despite the fact that Ziro in sentence (9) is not a subject. This is because Ziro is construed as the one reporting his feelings (Aikawa, 2002, pp. 170-1). The difference between the sentences is not structural; it is in whose internal thoughts are understood as being described.

Logophoricity is also relevant to explaining cases in which zibun does not refer to a subject antecedent. A logophoric individual, according to Aikawa (2002, p. 171) is someone other than the speaker whose attitudes, beliefs, feelings, or general states of consciousness are either being reported or are reflected in the linguistic context where the logophor appears (Clements, 1975). Aikawa (2002, p. 172) provides the following examples from Kuno (1973, pp. 309-310):
Again, despite the fact that these sentences are structurally similar (the only difference being *korosita* ‘killed’ versus *korosoo to sita* ‘tried to kill’), *zibun* can refer to *John* in one of them (11), but not in the other (10). By way of explanation, Kuno (1973) proposes that an antecedent for *zibun* must be aware of the situation at hand. Here, in the case where Mary merely tries to kill John, he can be aware of it; however, if she actually kills him, he cannot.

Empathy has also been argued to play a role in interpretations of *zibun* in the sense that the speaker of the sentence must be able to empathize with *zibun*’s antecedent (Kuno & Kaburaki, 1977). Aikawa (2002) explains that empathy “encodes whose point of view the speaker is taking” and indicates who the speaker identifies with (p. 188, note 26). For instance, sentences (12) and (13) below, from Aikawa (p. 173), differ only in the verb that is used, and yet they differ with respect to *zibun*’s possible antecedents.

(12)  *Tarooi-wa [Hanako-ga zibun-ni yatta] okane-o tukatte-simatta.*
Taro-TOP Hanako-NOM self-DAT gave money-ACC use-PERF
‘Taro has used the money that Hanako gave to self.’

(13)  *Tarooi-wa [Hanako-ga zibun-ni kureta] okane-o tukatte-simatta.*
Taro-TOP Hanako-NOM self-DAT gave money-ACC use-PERF
‘Taro has used the money that Hanako gave to self.’

In (12), the verb *yatta* (‘gave’) indicates that the speaker empathizes with Hanako, the giver, whereas in (13), the verb *kureta* (‘gave’) indicates that the speaker empathizes with Taro, the receiver. Thus, *zibun* can refer to Taro only in (13).

Another issue, according to arguments by Iida (1996), is that the antecedent associated with *zibun* indicates whose deictic perspective the speaker is taking, as a reference point, and the speaker can take only one of these at a time. Aikawa (2002, p. 174; based on Iida, p. 164) exemplifies this claim by presenting sentences with and without *zibun* which contain the deictic expression *migi* (‘on the right of’):

(14)  *Tarooi-wa Hanako-ni zitensya-o migi-ni ok-ase-ta.*
Taro-TOP Hanako-DAT bicycle-ACC right-to put-CAUSE-PAST
a. ‘Taro made Hanako put the bicycle to his right.’
b. ‘Taro made Hanako put the bicycle to her right.’

(15)  *Tarooi-wa Hanako-ni zibun-ni zitensya-o migi-ni ok-ase-ta.*
Taro-TOP Hanako-DAT self-GEN bicycle-ACC right-to put-CAUSE-PAST
a. ?? ‘Taro made Hanako put her bicycle to his right.’
b. ‘Taro made Hanako put her bicycle to her right.’

393
Imagining a situation in which Taro and Hanako are standing face to face, sentence (14), without *zibun*, is ambiguous; the bicycle could be placed either to Taro’s right or to Hanako’s right. However, if *zibun* is bound to Hanako, as in (15), the situation has to be interpreted as being described from Hanako’s perspective. Therefore, sentence (15) is unambiguous and can only mean that the bicycle was placed to Hanako’s right. Similarly, when a sentence contains two or more instances of *zibun*, they must refer to the same person (Iida, 1996, p. 80; originally from Howard & Niyekawa-Howard, 1976; cited by Aikawa, 2002, p. 176):

(16)  *Taro-wa [Hanako-ga zibun-no heya-de zibun-no sigoto-o site-ita to] itta.*
    Taro-TOP Hanako-NOM self-GEN room-in self-GEN job-ACC was-doingCOMP said
    a.  Taro said that Hanako was doing his work in his room.
    b.  Taro said that Hanako was doing her work in her room.
    c.  *Taro said that Hanako was doing his work in her room.
    d.  *Taro said that Hanako was doing her work in his room.

Of the issues reviewed above, that of verb choice may be one of the most important to reconsider in relation to some of the sentences used in the materials for this experiment (and in relation to some of the sentences used in experiments by generative SLA researchers). Citing examples such as sentences (17) and (18) below, Ueda (1986, p. 105) argues that an important factor in the acceptability of local binding of *zibun* is whether an abstract or concrete physical activity is being described (originally from Oshima, 1979, p. 425; cited by Aikawa, 2002, p. 183).

(17)  *Taro-ga zibun-o hihansita/semeta.*
    Taro-NOM self-ACC criticized/blamed
    ‘Taro criticized/blamed himself.’

(18)  *Taro-ga zibun-o nagutta/tataita/ketta.*
    Taro-NOM self-ACC hit/hit/kicked
    ‘Taro hit/hit/kicked himself.’

Sentences (17) and (18) are exactly parallel structurally, of course, and yet *zibun* can refer to the subject in (17), where abstract activities are involved, but cannot refer to the subject in (18), where concrete physical activities are being described. (As shown above in Appendices F1-F2, the verbs *tataku* (‘hit’) and *keru* (‘kick’) were employed in the materials for this experiment.)

Aikawa (2002) calls Ueda’s conclusion into question by presenting examples such as the following (pp. 184-5); however, she also notes that none of the considerations reviewed so far (e.g., awareness, perspective) can explain the distinction, either.

(19)  *Taro-ga zibun-no migi asi-o nagutta/tataita/ketta.*
    Taro-NOM self-GEN right leg-ACC hit/hit/kicked
    ‘Taro hit/hit/kicked his right leg.’

(20)  *Taro-ga [NP zibun-no kodomo]-o nagutta/tataita/ketta.*
    Taro-NOM self-GEN child-ACC hit/hit/kicked
    ‘Taro hit/hit/kicked his child’
Sentences (19-21) all involve the same concrete physical activities; the difference is that zibun appears either as a genitive or as the object of an embedded clause (as it often did in the sentences used in the present study). Aikawa (2002) explains this with reference to argument status: “[B]inding of argument zibun is affected by types of activities expressed by verbs, but binding of nonargument zibun is consistently grammatical” (p. 185).

Traditionally (e.g., Jespersen, 1933), reflexivization was considered to involve “the relationship between coarguments of a predicate” (Aikawa, 2002, p. 185); if a verb expressed a reflexive interpretation, a reflexive pronoun would be used. In Reinhart and Reuland’s (1993) well-known analysis of reflexivity, they make a distinction between morphologically complex ‘SELF’ anaphors, which can function as ‘reflexivizers’, and morphologically simple ‘SE’ anaphors, which cannot. If an anaphor is a reflexivizer, it can make a predicate reflexive (“impose identity between coarguments of a predicate”) even if the verb is not intrinsically reflexive (Aikawa, 2002, p. 180). For example, the anaphors zichzelf and zich in Dutch can be used to express that someone is ashamed of himself, but only the anaphor zichzelf, a reflexivizer, can be used to express that someone hates himself, as shown below in sentences (22-24), reproduced from Reinhart and Reuland (1993, pp. 661, 665, 666):

(22) Max, haat zichzelf.
‘Max hates SELF.’

(23) *Max, haat zich.
‘Max hates SE.’

‘Max is ashamed.’

The reason why (24) is grammatical whereas (23) is not is because the verb schaamt is intrinsically reflexive. The reason why (22) is grammatical whereas (23) is not is because the anaphor zichzelf can “induce” or “license” a reflexive interpretation of haat, but zich cannot.

Aikawa (1993) proposes that the binding behaviors of zibun (SE, a non-reflexivizer) and zibun-zisin (SELF, a reflexivizer) in Japanese can be explained in the same way, and that explains the difference between sentences (17) and (18) above: criticizing and blaming are listed in the lexicon as potentially reflexive, whereas hitting and kicking are not. Meanwhile, the uses of zibun in (19) and (20) are unproblematic because Taro and zibun are not co-arguments of the same predicate (“zibun-binding here has nothing to do with reflexivization of the predicates”) and, technically, non-argument zibun is not a reflexive (Aikawa, 2002, p. 185).

One final point to mention here is that a major reason behind diagramming causative structures in the way shown (see Figure 4.7) was linked directly to the notion of a subject requirement on zibun’s antecedents. Linguists often use interpretations of zibun as a diagnostic test of structural subjecthood in Japanese, and the fact that zibun can refer to a local antecedent marked accusative or dative in a causative has been taken to imply, for one thing,
that that antecedent must therefore be a structural subject, and, for another, that certain types of causatives must be biclausal. As Aikawa (2002) points out, if *zibun* does not necessarily refer to grammatical subjects, then analyses of such structures as biclausal might be called into question. For the present experiment, the upshot is that, despite having been based on standard analyses in the field of Japanese theoretical linguistics (e.g., Tsujimura, 2007), the diagrams of causative structures might not have represented mental reality, whatever that is.

A strict c-command requirement may also be questionable. Kim, Montrul, and Yoon (2010) have pointed out that in Korean and Chinese (but not in English), “in configurations called ‘sub-command’ (Tang, 1989),” if an antecedent is the possessor of an inanimate noun which c-commands a reflexive, that possessor can bind the reflexive despite not c-commanding it (pp. 77-8). For example (in Korean):

(25)  *
\[ \text{Silvia-uy casonsím-i cakí-i-lul koylóphi-n-ta} \]
\[ \text{Silvia-GEN pride-NOM self-ACC torture-PRES-DECL} \]
\[ \text{‘Silvia’s pride tortures herself (= Silvia)’ [possible in Korean, impossible in English]} \]

(26)  *
\[ \text{Silvia-uy tongsayng-i cakí-i-lul koylóphi-n-ta} \]
\[ \text{Silvia-GEN brother-NOM self-ACC torture-PRES-DECL} \]
\[ \text{‘Silvia’s brother tortures herself (= Silvia)’ [impossible in both Korean and English]} \]

One point in reviewing these controversies over *zibun* is to suggest that if researchers investigating the SLA of reflexives assume that conditions on their interpretation are purely structural, we may very well draw misguided conclusions. The present experiment was set up to focus on c-commanding subjects (local or long-distance) as possible antecedents for *zibun*, but that does not mean that the participants had not had experience with other types of antecedents outside of the experiment, and it is important to bear in mind that explicitly structure-oriented information was provided in only one of the treatment conditions (i.e., when tree diagrams were given as feedback). It cannot be assumed that participants in all of the groups were entertaining structure-related hypotheses, but simply with varying degrees of accuracy. Some might have been contemplating pragmatic factors involving perspective or empathy, for instance, and may have shown improvement to the extent that hypotheses of that nature were also accurate. Such a possibility is discussed in Chapter 6 in relation to the performance of participants in the Right/Wrong-Feedback condition, considering some unexpected differences across the groups in the relationships between individual differences and test performance. The introspective data which some of the participants produced in their concurrent think-alouds during the treatments will be especially valuable here.
Appendix G: Motivation and reflection questionnaires

Each participant filled out a motivation questionnaire at the very beginning of the experiment, another type of motivation questionnaire following the pre-test and each treatment session, and a final type of motivation questionnaire following the immediate post-test. These questionnaires were designed in collaboration with Dr. Heather Weger to probe the learners’ beliefs regarding their process-oriented motivation, which Dörnyei (2003) argued to be comprised of three stages: pre-actional, actional, and post-actional.

The pre-actional stage, a goal-setting stage, is theorized to be influenced by multiple affective variables, including learners’ attitudes toward the L2, values associated with learning, and expectancy of success. Accordingly, the pre-actional motivation questionnaire (Appendix G1) consisted of two parts: (1) a section asking about the learners’ reasons for participating in the research project and probing for causal attributions of learning success (Hsieh & Kang, 2010; Julkunen, 1989), and (2) a section designed to tap into the learners’ perceptions of their attitudes toward the L2 by asking about their “possible selves” (Dörnyei, 2009). The participants completed the pre-actional questionnaire on the first day of the experiment, after the informed consent process, but before any of the other activities.

The next stage of process-oriented motivation, the actional stage, is the stage in which planned actions are executed. It is theorized to be influenced by such variables as the quality of a learning activity, learners’ appraisals of their achievement, and their self-regulated engagement with the activity. Following the pre-test, the actional questionnaire (Appendix G2) contained 10 Likert-style items which asked the participants to evaluate their experience with regard to the activity’s level of difficulty; their enjoyment, effort, and concentration (Julkunen, 1989); their level of comfort with computers (Nagata, 1995); and the projected usefulness of similar activities in the future. Following each treatment, the actional questionnaire (Appendix G3) contained 3 additional items designed to probe whether the participants had understood the information provided by the computer, whether they had considered it helpful, and whether they felt more information would have been helpful.

Finally, the post-actional stage of motivation is the stage in which learners reflect on their past experiences. It is theorized to be influenced by such variables as attributional factors and self-concept beliefs. The participants completed the post-actional questionnaire (Appendix G4) immediately after the post-test. Like the pre-actional questionnaire, it consisted of two sections: (1) a section probing for the participants’ perceptions of learning and attributions of learning success (Hsieh & Kang, 2010; Julkunen, 1989), and (2) a section designed to tap into the learners’ attitudes toward the L2 by asking about their “possible selves” (Dörnyei, 2009).
Appendix G1
Motivation questionnaire—Preactional

**Project:** Learners of Japanese

**Participant ID:** ______________

**DIRECTIONS:** We would like to ask you to help us by answering the following questions about foreign language learning. This is NOT a test, so there are no “right” or “wrong” answers. We are interested in your personal opinion.

**Part 1: Participating in this research project.** Next to each statement below please indicate your opinion by putting an “X” in the box that best indicates how much you agree or disagree with the statement.

<table>
<thead>
<tr>
<th><strong>Example:</strong> I like to eat pizza.</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I want to participate in this research project because I think I can learn something about Japanese.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. I want to participate in this research project because I like to do activities on the computer.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. I want to participate in this research project because I think computer activities are a good way to learn Japanese.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. I want to participate in this research project because my teacher encouraged me to volunteer.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. I want to participate in this research project because it will help the research team.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Please explain any other reason why you agreed to participate in this research project.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. When I am learning Japanese, I like to know the rules.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. When I am learning Japanese, I like to figure things out on my own.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. I am certain I can learn something about Japanese while I am doing this project because I am a good language learner.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. I am certain I can learn something about Japanese while I am doing this project because I am good at learning Japanese grammar.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. I am certain I can learn something about Japanese while I am doing this project because I am a hard worker.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. I am certain I can learn something about Japanese while I am doing this project because the activities will be interesting.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. I am certain I can learn something about Japanese while I am doing this project because I will enjoy doing the activities.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. I am certain I can learn something about Japanese while I am doing this project because I trust that the project designers are good.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Please explain any other reason why you think you can learn something about Japanese while doing this project.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Part 2: Possible Selves in Japanese:** Probably everyone thinks about the future to some extent. When doing so, we often think about the kinds of experiences that are in store for us. Listed below are a number of possibilities of “future selves” in Japanese that have been provided by other people. We are interested in what “possible selves” (both positive and negative) you have considered for yourself.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>This “possible self” actually characterizes me right now. (circle one) YES / NO</th>
<th>This “possible self” characterizes me in my future. (circle one) YES / NO</th>
<th>Is this desired or not desired in your future? 1=undesirable 2=somewhat undesirable 3=neutral 4=somewhat desirable 5=desirable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Understand the views of Japanese people</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Think like the Japanese people</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Act like the Japanese people</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Understand Japanese literature</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Appreciate Japanese art and literature</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Watch movies &amp;/or TV shows in Japanese</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Read newspapers &amp;/or magazines in Japanese</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Read/listen to things from the internet in Japanese</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Feel at ease with Japanese people</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Have friendships with Japanese people</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Travel to Japan for vacations</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Travel to Japan for work</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Work at a job in Japan</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Work at a job in my home country (the U.S.) using Japanese</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Give presentations or attend meetings using Japanese for my job</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Talk to customers or co-workers in Japanese for my job</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Write in Japanese for my job</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Read materials in Japanese for my job</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Be a knowledgeable person</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Be a cultured person</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Feel respected because I speak Japanese</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Enjoy speaking Japanese</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Want to learn many languages</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Participate freely in activities of other cultural groups</td>
<td>YES / NO</td>
<td>YES / NO</td>
<td></td>
</tr>
</tbody>
</table>
Appendix G2
Motivation questionnaire—Actional #1

Project: Learners of Japanese

**Participant ID:**

**DIRECTIONS:** We would like to ask you to help us by answering the following questions about foreign language learning. This is NOT a test, so there are no “right” or “wrong” answers. We are interested in your **personal opinion**.

Next to each statement below please indicate your opinion by putting an “x” in the box that best indicates how much you agree or disagree with the statement.

<table>
<thead>
<tr>
<th>Example: I like to eat pizza.</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I enjoyed the activity.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. The activity was easy.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. I was comfortable interacting with the computer.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. I was able to concentrate when I was doing this activity.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. I had enough time to do the activity.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. The instructions for the activity were clear.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. I put my best effort into doing this activity <strong>today</strong>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. I will put my best effort into doing the next activity.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. I learned something about Japanese when I was doing this activity <strong>today</strong>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Doing another activity like this one will be helpful for my language learning.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Is there anything else you would like to tell us about today’s activity?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix G3  
Motivation questionnaire—Actional #2

<table>
<thead>
<tr>
<th>Example: I like to eat pizza.</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I enjoyed the activity.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. The activity was easy.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. I was comfortable interacting with the computer.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. I was able to concentrate when I was doing this activity.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. I had enough time to do the activity.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. The instructions for the activity were clear.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. I understood the information that the computer activity gave me.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. The information that the computer activity gave me was helpful.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. More information during the computer activity would have been helpful.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. I put my best effort into doing this activity <em>today</em>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. I will put my best effort into doing the <em>next</em> activity.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. I learned something about Japanese when I was doing this activity <em>today</em>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Doing another activity like this one will be helpful for my language learning.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Is there anything else you would like to tell us about today's activity?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix G4
Motivation questionnaire—Post-actional

DIRECTIONS: We would like to ask you to help us by answering the following questions about foreign language learning. This is NOT a test, so there are no “right” or “wrong” answers. We are interested in your personal opinion. Next to each statement below please indicate your opinion by putting an “x” in the box that best indicates how much you agree or disagree with the statement.

**QUESTION 1**
1. Think about all the computer activities you did in Japanese during this project. Do you think you learned something about Japanese when you were doing these computer activities?

**If no →**
If you said “NO,” answer these questions.

<table>
<thead>
<tr>
<th>Example: I like to eat pizza.</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. I did NOT learn something about Japanese when I was doing these activities because I am NOT a good language learner.</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1b. I did NOT learn something about Japanese when I was doing these activities because I did NOT work hard.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1c. I did NOT learn something about Japanese when I was doing these activities because I was NOT lucky.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1d. I did NOT learn something about Japanese when I was doing these activities because the activity was NOT interesting.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1e. I did NOT learn something about Japanese when I was doing these activities because I did NOT enjoy doing the activity.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1f. I did NOT learn something about Japanese when I was doing these activities because the computer activity did NOT provide helpful information.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1g. Please explain any other reason why you did NOT learn something about Japanese when you were doing these activities:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**If yes →**
If you said “YES,” answer these questions.

<table>
<thead>
<tr>
<th>Example: I like to eat pizza.</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1h. I learned something about Japanese when I was doing these activities because I am a good language learner.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1i. I learned something about Japanese when I was doing these activities because I worked hard.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1j. I learned something about Japanese when I was doing these activities because I was lucky.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1k. I learned something about Japanese when I was doing these activities because the activity was interesting.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1l. I learned something about Japanese when I was doing these activities because I enjoyed doing the activity.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1m. I learned something about Japanese when I was doing these activities because the computer activity provided helpful information.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1n. Please explain any other reason why you learned something about Japanese when you were doing these activities:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. Do you want to study this grammar structure more?  

**NO** (answer questions 2a-2e; skip 2f-2j)  
**YES** (answer questions 2f-2j; skip 2a-2e)

### If no →

<table>
<thead>
<tr>
<th>IF YOU SAID “NO,” answer these questions.</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Example:</strong> I like to eat pizza.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2a. I do NOT want to study this grammar structure more because I understand it.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2b. I do NOT want to study this grammar structure more because it is not important.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2c. I do NOT want to study this grammar structure more because it is difficult.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2d. I do NOT want to study this grammar structure more because it is not interesting.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2e. Please explain any other reason why you do NOT want to study this grammar structure more:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### If yes →

<table>
<thead>
<tr>
<th>IF YOU SAID “YES,” answer these questions.</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Example:</strong> I like to eat pizza.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2f. I want to study this grammar structure more because I do not understand it yet.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2g. I want to study this grammar structure more because it is important.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2h. I want to study this grammar structure more because it is easy.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2i. I want to study this grammar structure more because it is interesting.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2j. Please explain any other reason why you want to study this grammar structure more:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 3. Do you want to use more computer activities to study other Japanese grammar structures?

<table>
<thead>
<tr>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>YES</td>
</tr>
</tbody>
</table>

**If no →**

**IF YOU SAID “NO,” answer these questions.**

<table>
<thead>
<tr>
<th>Example: I like to eat pizza.</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3a. I do NOT want to use computer activities to study other Japanese grammar structures because I do not like to do activities on the computer.

3b. I do NOT want to use computer activities to study other Japanese grammar structures because I do not think that computer activities are a good way to learn Japanese.

3c. I do NOT want to use computer activities to study other Japanese grammar structures because I do not trust that the program designers are good.

3d. Please explain any other reason why you do NOT want to use computer activities to study other Japanese grammar structures:

**If yes →**

**IF YOU SAID “YES,” answer these questions.**

<table>
<thead>
<tr>
<th>Example: I like to eat pizza.</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3e. I want to use computer activities to study other Japanese grammar structures because I like to do activities on the computer.

3f. I want to use computer activities to study other Japanese grammar structures because I think that computer activities are a good way to learn Japanese.

3g. I want to use computer activities to study other Japanese grammar structures because I trust that the computer program designers are good.

3h. Please explain any other reason why you want to use computer activities to study other Japanese grammar structures:
Part 3: Possible Selves in Japanese: Probably everyone thinks about the future to some extent. When doing so, we often think about the kinds of experiences that are in store for us. Listed below are a number of possibilities of “future selves” in Japanese that have been provided by other people. We are interested in what “possible selves” (both positive and negative) you have considered for yourself.

<table>
<thead>
<tr>
<th></th>
<th>This “possible self” actually characterizes me right now. (circle one) YES / NO</th>
<th>This “possible self” characterizes me in my future. (circle one) YES / NO</th>
<th>Is this desired or not desired in your future? 1=undesirable 2= somewhat undesirable 3=neutral 4= somewhat desirable 5=desirable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Understand the views of Japanese people</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>2</td>
<td>Think like the Japanese people</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>3</td>
<td>Act like the Japanese people</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>4</td>
<td>Understand Japanese literature</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>5</td>
<td>Appreciate Japanese art and literature</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>6</td>
<td>Watch movies &amp;/or TV shows in Japanese</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>7</td>
<td>Read newspapers &amp;/or magazines in Japanese</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>8</td>
<td>Read/listen to things from the internet in Japanese</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>9</td>
<td>Feel at ease with Japanese people</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>10</td>
<td>Have friendships with Japanese people</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>11</td>
<td>Travel to Japan for vacations</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>12</td>
<td>Travel to Japan for work</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>13</td>
<td>Work at a job in Japan</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>14</td>
<td>Work at a job in my home country (the U.S.) using Japanese</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>15</td>
<td>Give presentations or attend meetings using Japanese for my job</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>16</td>
<td>Talk to customers or co-workers in Japanese for my job</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>17</td>
<td>Write in Japanese for my job</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>18</td>
<td>Read materials in Japanese for my job</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>19</td>
<td>Be a knowledgeable person</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>20</td>
<td>Be a cultured person</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>21</td>
<td>Feel respected because I speak Japanese</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>22</td>
<td>Enjoy speaking Japanese</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>23</td>
<td>Want to learn many languages</td>
<td>YES / NO</td>
<td>YES / NO</td>
</tr>
<tr>
<td>24</td>
<td>Participate freely in activities of other cultural groups</td>
<td>YES/ NO</td>
<td>YES / NO</td>
</tr>
</tbody>
</table>
Appendix G5
Reflection questionnaire

Researchers interested in learners’ awareness of linguistic targets under more versus less informative treatment conditions often use post-task debriefing questionnaires as one means of exploring what the learners have noticed, whether they were looking for rules, whether they can state any rules, and what they have understood the purpose of the experiment to be (e.g., Leow, 2000; Robinson, 1997a, 1997b; Sachs & Suh, 2007). In the current study, such a questionnaire was administered to both the think-aloud participants and the non-think-aloud participants in order to gain an additional perspective on the learners’ conscious appraisals of what they had been through and what they had learned. The questions were as follows:

1. At the very beginning of this study, as you probably remember, I explained its rationale in fairly general terms (i.e., to explore how different sorts of Japanese learning activities may function differently for people with different learning styles and orientations). In the course of participating in the experiment, do you think you’ve gotten a more specific sense of what the study is about?

2. How might you describe the approach you took to completing the Japanese learning activities?

3. Please elaborate on your answer to #2 by giving specific examples if you can. (Don’t worry if you can provide just a partial or tentative description – any information you can share on your thought processes will be very helpful!)

4. Have you noticed anything interesting about how Japanese works that you didn’t know before?

5. From the time you started participating until now, have you talked with anyone else about the computer activities you’ve been doing? Have you asked anyone about particular language points or looked up anything that seemed relevant? If so, please explain.

6. Any other comments?
Appendix H
Tests of working memory capacity

Listening-span test

The listening-span test employed in this study was based on Daneman and Carpenter’s (1980) reading-span task. It was developed and validated at Georgetown University as part of The Latin Project by Dr. Cristina Sanz, Dr. Harriet Wood Bowden, and Dr. Catherine Stafford, with the assistance of William Garr. It was accessed with the permission of Dr. Sanz. Similar versions have been used in several studies of SLA (e.g., Lado, 2008; Mackey et al., 2002; Mackey et al., 2010; Mackey & Sachs, in press; Winke, 2005).

The test involved listening to sets of unrelated sentences in English, judging whether or not each was grammatical and whether or not it made sense, and remembering the final word of each sentence in the set for subsequent recall. Over the course of the test, the number of sentences in each set was gradually increased, thereby increasing the difficulty of keeping the final words in memory while processing the other sentences for grammaticality and meaning.

It was administered individually via computer to each participant, who listened to the sentences through headphones, marked their judgments on a paper answer sheet (H1), spoke the recall words aloud at the end of each set (recording them with a handheld digital recorder or onto the computer), and clicked to advance from set to set at their own pace. The test included 48 sentences, pre-recorded by a female native speaker of English. They were presented in sets of 3, 4, or 5 sentences each, with a 2-second pause after each sentence in a set. There were a total of 12 sets, with 4 sets at each set size. Most of the recall items were non-compound nouns (15 monosyllabic, 25 disyllabic, and 7 trisyllabic, plus one 5-syllable word [organization]), and none of the words in a set were judged by the test developers to be semantically related to one another.

As part of the instructions before beginning the test, the participants completed 3 practice items (reproduced here for illustrative purposes):

1. *Working at the feet is a great job for Tom, but standing all day is beginning to wear on his bank* (nonsense, grammatical)
2. *My older sister’s furniture were really small, but she has a lot of apartment and no space for storage* (nonsense, ungrammatical)
3. *She never goes camping because she doesn’t like have to use bathrooms without running water* (sense, ungrammatical)

The scoring involved awarding 1 point for each correctly recalled word (ignoring differences in plurality, such as *bank* vs. *banks*), 1 point for each correct sense judgment, and 1 point for each grammaticality judgment. After multiplying the recall score by 2 to balance the contributions of storage and processing, the total possible score was 192 (96 for processing, including 48 for sense and 48 for grammaticality, and 96 [48 x 2] for storage).
Operation-span test

The operation-span test employed in this study was adapted from a Korean-based version provided by Jaemyung Goo, a fellow Ph.D. candidate at Georgetown University. That version was, in turn, a modified version of Turner and Engle’s (1989) operation-span test. It involved holding series of letters in memory while performing simple mathematical calculations. There were 2 practice sets, each of which included 3 equations (and letters to be recalled), followed by 10 test sets composed of 4, 5, 6, 7, or 8 equations (and letters) each, for a total of 60 items. There were 2 sets at each level, gradually increasing in size over the course of the test.

Each equation involved 2 operations: either multiplication or division, followed by either addition or subtraction. For example, one item was \((2 \times 4) - 3\) and another was \((6 / 2) + 5\). Only single-digit numbers were used. Half of the equations were correct, whereas the other half were incorrect, and the operations involved in them (i.e., addition, subtraction, multiplication, and division) were also balanced. The letters which the participants had to recall were taken from research on the implicit learning of artificial grammars: letter sets 1 (SCVTXP), 2 (FDMQJH), and 3 (NKZLYB) from Mathews, Buss, Stanley, Blanchard-Fields, Cho, and Druhan (1989); MVXTR from Perruchet and Pacteau (1990), and TVXPS from Reber (1967). Each of these 18 letters \((B C F H J K L M N P Q R S T V X Y Z)\) was used 3 or 4 times over the course of the test, never in the same set.

The test was administered via Microsoft PowerPoint, using custom animation features to cause the slides to advance automatically. All of the slides had plain white backgrounds with black text. The equations were presented in 54-point font and the letters to be recalled were presented in 72-point font. The participants marked their responses on a two-sided paper answer sheet (H2). They were given 3 seconds to solve each equation, 3 seconds to mark on the front side of the answer sheet whether the provided answer was correct or incorrect, and 1 second to view the letter they were supposed to recall. At the end of each set, question marks appeared on the screen (???), and the participants flipped over to the back side of their answer sheet in order to write down, in order, all of the letters they recalled for that set. The number of seconds allotted for this depended on the number of letters to be recalled; specifically, there was a base of 2 seconds plus 1 additional for each letter in the set (i.e., 6, 7, 8, 9, or 10 seconds).

In scoring the operation-span test, 1 point was awarded for each correct equation judgment (= 60 points total for the processing component), and 1 point was awarded for each correctly recalled letter written in the right order, relatively speaking (= 60 points total for the storage component). The participants were given a half point for each recalled letter which was simply misplaced, and a half point was subtracted for any extra letter they included beyond those which could be considered wrong-letter substitutions. (In other words, if a wrong letter could be interpreted as simply standing in for another letter that was supposed to be in that position, it was simply counted as wrong and 0 points were awarded or subtracted. However, if a letter was written again after already having been included in the string, or if an extra letter was tacked onto the end or intervened between letters which were supposed to be adjacent, a half point was subtracted.)
**Appendix H1**

**Listening-span test answer sheet**

(page 1)

Listen to the sentences and make two judgments: does the sentence make sense and is it grammatical?

**Practice**

<table>
<thead>
<tr>
<th>Does it make sense?</th>
<th>Is it grammatical?</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Yes</td>
<td>No</td>
</tr>
<tr>
<td>b. Yes</td>
<td>No</td>
</tr>
<tr>
<td>c. Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

**Actual task**

<table>
<thead>
<tr>
<th>Does it make sense?</th>
<th>Is it grammatical?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Yes</td>
<td>No</td>
</tr>
<tr>
<td>2. Yes</td>
<td>No</td>
</tr>
<tr>
<td>3. Yes</td>
<td>No</td>
</tr>
<tr>
<td>4. Yes</td>
<td>No</td>
</tr>
<tr>
<td>5. Yes</td>
<td>No</td>
</tr>
<tr>
<td>6. Yes</td>
<td>No</td>
</tr>
<tr>
<td>7. Yes</td>
<td>No</td>
</tr>
<tr>
<td>8. Yes</td>
<td>No</td>
</tr>
<tr>
<td>9. Yes</td>
<td>No</td>
</tr>
<tr>
<td>10. Yes</td>
<td>No</td>
</tr>
<tr>
<td>11. Yes</td>
<td>No</td>
</tr>
<tr>
<td>12. Yes</td>
<td>No</td>
</tr>
<tr>
<td>13. Yes</td>
<td>No</td>
</tr>
<tr>
<td>14. Yes</td>
<td>No</td>
</tr>
<tr>
<td>15. Yes</td>
<td>No</td>
</tr>
<tr>
<td>16. Yes</td>
<td>No</td>
</tr>
<tr>
<td>17. Yes</td>
<td>No</td>
</tr>
<tr>
<td>18. Yes</td>
<td>No</td>
</tr>
<tr>
<td>19. Yes</td>
<td>No</td>
</tr>
<tr>
<td>20. Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
Appendix H2
Operation-span test answer sheet

Operation span answer sheet
SIDE 1

Was the solution to the equation correct?

PRACTICE

<table>
<thead>
<tr>
<th></th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ACTUAL TASK

<table>
<thead>
<tr>
<th></th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Which letters did you see along with the most recent set of equations?

**PRACTICE**

<table>
<thead>
<tr>
<th>Set 1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 2</td>
<td></td>
</tr>
</tbody>
</table>

**ACTUAL TASK**

<table>
<thead>
<tr>
<th>Set 1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 2</td>
<td></td>
</tr>
<tr>
<td>Set 3</td>
<td></td>
</tr>
<tr>
<td>Set 4</td>
<td></td>
</tr>
<tr>
<td>Set 5</td>
<td></td>
</tr>
<tr>
<td>Set 6</td>
<td></td>
</tr>
<tr>
<td>Set 7</td>
<td></td>
</tr>
<tr>
<td>Set 8</td>
<td></td>
</tr>
<tr>
<td>Set 9</td>
<td></td>
</tr>
<tr>
<td>Set 10</td>
<td></td>
</tr>
</tbody>
</table>
Appendix H3
Instructions for the operation-span test

SLIDE 1:
This activity involves holding series of letters (e.g., X V T) in memory while you perform simple mathematical calculations (e.g., \((2 \times 1) + 2\)).

All in all, you will complete 60 simple math problems. Over the course of the activity, you will be prompted 10 times to recall sequences of letters that you have recently seen.

The following slides will explain the procedure in more detail.

SLIDE 2:
First, you will see a math problem, like \((2 \times 1) + 2\).

Every problem will involve a simple multiplication or division, followed by a simple addition or subtraction. Please solve the equation in your head.

After 3 seconds, the screen will advance automatically, and you will see a single number (e.g., 4) along with the question “Yes or No?”

Decide whether this number matches the answer to the equation you have just solved. If it matches, circle “Yes” on your answer sheet. If it does not match, circle “No”.

This screen will also advance automatically after 3 seconds.

SLIDE 3:
On the screen after that, you will see a letter (e.g., V). This will be shown for 1 second and then advance to the next screen automatically. Hold each of the letters you see in memory because you will be asked to recall them.

Once in a while, the screen will contain three question marks (???). When this happens, try to recall all of the letters you have seen for the most recent set of equations. Please write these letters in order on the back of your answer sheet.

In other words, every time you see “???” write down all the letters you have seen since the previous time you saw “???”.

Don’t forget to write the letters in their order of appearance. If you don’t remember a letter, just put a blank line where the letter was supposed to go.
SLIDE 4:

To summarize:

1. **Math problem**: Solve the equation in your head (in 3 seconds).
   
   \[(6 / 2) − 1 = ?\]

2. **Judgment**: Decide (again, in 3 seconds) whether the number shown on the screen matches the answer to the equation you have just solved.
   
   \[3\]
   
   Yes or No?
   
   (For this problem, of course, you would mark “No” on your answer sheet.)

3. **Letter**: Look at the letter presented on the screen and remember it. (This will be shown for 1 second.)
   
   \[V\]

4. **Next step**: You will see either (a) another math problem, or (b) “???”.

   If you see a math problem, keep going and follow the same steps as above, continuing to keep in mind the letters you are seeing.

   If you see a slide with “???” , recall all the letters you have seen since the last time you saw “???” and write them down, in order, on the back of your answer sheet.

SLIDE 5:

Just FYI:

The [equation]-[judgment]-[letter] sequences will be presented in sets of 3, 4, 5, 6, 7, or 8 in a row.

This means that when you are prompted with “???” , you will have to recall from between 3-8 letters at a time. You’ll have between 5-10 seconds to recall the letters, depending on how many letters you need to recall.

The “???” slides will also advance automatically when the time is up.
Appendix I
Test of sensitivity to ambiguity and familiarity with metalinguistic terms

Participant ID: ____________

Part 1: Fun with ambiguous sentences!

Sometimes, a single sentence can have more than one possible meaning. For example, think about the exclamation “We need more intelligent teachers!”

It could be interpreted as meaning either:
   1. We need teachers who are more intelligent, or
   2. We need a greater number of intelligent teachers

Many (but not all) of the sentences you are about to see can be interpreted in more than one way. Some of the meanings may be more straightforwardly obvious than others, so please try your best to see the different possibilities. At the same time, don’t worry too much if you only see one meaning—some of the sentences below actually are unambiguous, with only one possible meaning. That’s important, too.

Basically, just trust yourself and go with the meaning(s) that truly seem possible for each sentence. 😊

Please try to paraphrase the meaning(s) of each sentence below in a very clear, unambiguous way. If you get only one reading for a sentence, just leave the additional lines blank.

Here we go!

1. Joe loves his mother, and so does Jim.


3. The podiatrist envied the shoe salesman with two assistants who loved feet.

4. The passenger saw the taxi driver looking at himself in the mirror.

6. It didn’t snow on more than 2 days.

7. The mother made her daughter listen to herself.

8. Picasso discussed painting with Dali.

9. The professor who saw the student cheating blamed himself.

10. A girl kissed every boy.

11. The intern recognized the patient of the nurse who fell asleep at the hospital.

12. The pitching coach made the baseball player watch himself.

13. The safari guide saw a birdwatcher with binoculars.
14. All of Angelina’s kids aren’t asleep.

15. The amnesic asked the psychologist some questions about himself.

16. The homecoming queen that the mean cheerleader punched suspected that the valedictorian hated herself.

Part 2: Fun with grammatical terminology!

For each grammar term given in the list on the next page, please find an example in the paragraph below and write it on the appropriate line. In each case, include a bit of the surrounding context and circle only the relevant item(s). The first one has been done for you. (By the way, don’t worry if you aren’t familiar with all of these terms—it’s normal! Not everyone studies this stuff explicitly in school!)

I’ll always remember the day after the aliens ate my co-workers’ brains. My boss Ethel, a sweet-looking old grandmother-type, had started bringing in desserts that were giving everyone severe indigestion. Employees who had been on the job for a while knew that her desserts (despite looking delicious!) were completely rancid, so they politely declined. Luckily, soon after I was hired, I found out who was in the know and made them give me the scoop on everything. However, Ethel, who was always looking for people to try her repulsive concoctions, cajoled some other new employees into trying them. They couldn’t help themselves! What they didn’t know was that those desserts tasted so bad because they had alien spores in them. One day, of course, the aliens hatched, burst through my co-workers’ bellybuttons, and started feasting on their brains. An eerie stillness came over the office after the aliens had finished eating. That’s why I’ll never forget it—it was nice and quiet for once!

Example: Subject  **You can imagine**
1. Direct object

2. Indirect object

3. Intransitive verb

4. Transitive verb

5. Ditransitive verb

6. Possessive pronoun

7. Reflexive pronoun

8. Antecedent of a personal pronoun

9. Subordinating conjunction

10. Adjectival subordinate clause

11. Adverbial subordinate clause

12. Restrictive relative clause

13. Nonrestrictive relative clause

14. Interrogative nominal clause

15. Causative

16. Appositive
Appendix J
Visual Patterns Test sample answer sheet

13(6) 16(7)

14(6) 17(7)

15(6) 18(7)
Appendix K
Background questionnaire

Participant ID: ____________

Name: ________________________  Age: ______  Male or female? __________

What is/are your native language(s)? ______________________________________

What Japanese course(s) are you enrolled in this semester? _______________________

How long have you been studying Japanese? __________________________________

Please explain where and with what intensity you have studied Japanese (for example, 2 years of daily classes in high school, 2 semesters of intensive university classes, a summer at Middlebury...).
__________________________________________________________________________

Why are you studying Japanese? _____________________________________________
__________________________________________________________________________

Have you spent any time living or traveling in Japan? ____________________________

About how many hours a week do you use Japanese outside the classroom and what do you use it for? ________________________________
__________________________________________________________________________

Do you know any other languages besides English and Japanese? Which languages and with what proficiency levels and skills?
__________________________________________________________________________
__________________________________________________________________________
__________________________________________________________________________

How did you learn those other languages? _____________________________________
__________________________________________________________________________

What is your major? _________________________________________________________

Have you ever taken a course in linguistics? if so, which? _______________________

Have you ever learned how to diagram sentences? If so, when and using what system? ______________
__________________________________________________________________________

Was diagramming sentences relatively easy for you or pretty difficult? (Please elaborate.) ______________
__________________________________________________________________________

Would you say that you like studying grammar, or is it not really your thing? (Please elaborate.)
__________________________________________________________________________

Thanks a lot! This information will be very helpful!

419
Appendix L
Reactivity of thinking aloud

This appendix presents tables and graphs associated with the analyses conducted to determine whether thinking aloud during the treatments affected the amount of improvement participants demonstrated over the course of the tests—in other words, whether thinking aloud was reactive. A repeated-measures analysis of variance (RM ANOVA) was used to answer this question, comparing learners in the Think-Aloud (TA) condition against learners in the Non-Think-Aloud (Non-TA) condition. Two participants from the Trees-Feedback condition (P7, in the TA group, and P59, in the Non-TA group) were excluded due to their failure to complete one of the tests each.

Figure L.1 and Tables L.1-L.3 display information relevant to checking the assumptions of normality of distributions, homogeneity of variances, and sphericity. The goodness-of-fit tests indicate that there might be departures from normality for the TA group on the pre-test (Kolmogorov-Smirnov) and the Non-TA group on the delayed post-test (Shapiro-Wilk). The skewness and kurtosis ratios are within the acceptable range of +/-2 for both groups on all tests, however, and although the boxplots suggest a slight amount of negative skew for the TA group on the pre-test and a slight amount of positive skew for the Non-TA group on the delayed post-test, these do not jump out as blatant violations. Levene’s tests of homogeneity produced no statistically significant results, and the standard deviations in Table L.1 are all within about 2 percentage points of each other. The interquartile and overall ranges shown in the boxplots do not seem to differ much, except possibly in the case of the post-test, where the TA group’s IQR looks slightly wider. Mauchly’s test of sphericity was statistically significant; thus, the Greenhouse-Geisser correction was used in the ANOVA.

An affirmative answer to the question of whether the TA and Non-TA groups differed from each other in the amount of improvement they showed would require that the RM ANOVA produce a statistically significant Time-by-Think-Aloud interaction effect. As discussed in the main text of Chapter 4 and displayed in Table L.4, this was not found. Table L.1 and Figures L.1 and L.2 show that the mean test scores of learners who produced think-aloud protocols during the treatments were always within 4 percentage points of the mean test scores of those who remained silent.
Figure L.1. Boxplots of pre-, post-, and delayed post-test scores for the TA and Non-TA groups

Figure L.2. Comparison of TA and Non-TA groups with regard to test performance
Table L.1. Descriptive statistics of pre-, post-, and delayed post-test scores across think-aloud conditions

<table>
<thead>
<tr>
<th>Test</th>
<th>Group</th>
<th>Mean</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Std. Dev.</th>
<th>Variance</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lower</td>
<td>Upper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-test</td>
<td>Non-Think-Aloud</td>
<td>60.555</td>
<td>38.57</td>
<td>40.00</td>
<td>78.57</td>
<td>10.253</td>
<td>105.131</td>
<td>57.231 - 63.879</td>
</tr>
<tr>
<td></td>
<td>Think-Aloud</td>
<td>57.874</td>
<td>44.53</td>
<td>32.14</td>
<td>76.67</td>
<td>11.579</td>
<td>134.063</td>
<td>54.120 - 61.627</td>
</tr>
<tr>
<td>Post-test</td>
<td>Non-Think-Aloud</td>
<td>67.279</td>
<td>56.67</td>
<td>43.33</td>
<td>100.00</td>
<td>13.171</td>
<td>173.476</td>
<td>63.010 - 71.549</td>
</tr>
<tr>
<td></td>
<td>Think-Aloud</td>
<td>68.004</td>
<td>59.26</td>
<td>40.74</td>
<td>100.00</td>
<td>15.352</td>
<td>235.695</td>
<td>63.028 - 72.981</td>
</tr>
<tr>
<td>Delayed post-test</td>
<td>Non-Think-Aloud</td>
<td>63.596</td>
<td>42.86</td>
<td>46.43</td>
<td>89.29</td>
<td>11.799</td>
<td>139.226</td>
<td>59.771 - 67.421</td>
</tr>
<tr>
<td></td>
<td>Think-Aloud</td>
<td>67.528</td>
<td>56.67</td>
<td>43.33</td>
<td>100.00</td>
<td>13.064</td>
<td>170.662</td>
<td>63.293 - 71.763</td>
</tr>
</tbody>
</table>
Table L.2. Goodness-of-fit tests to check normality of distributions of pre, post, and delayed scores across think-aloud conditions

<table>
<thead>
<tr>
<th>Test</th>
<th>Group</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Kolmogorov-Smirnov&lt;sup&gt;§&lt;/sup&gt;</th>
<th>Shapiro-Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n=39 per group)</td>
<td>Skew</td>
<td>Std. Error</td>
<td>Ratio</td>
<td>Kurtosis</td>
</tr>
<tr>
<td>Pre-test</td>
<td>Non-TA</td>
<td>-.165</td>
<td>.378</td>
<td>-0.437</td>
<td>-.617</td>
</tr>
<tr>
<td></td>
<td>Think-Aloud</td>
<td>-.545</td>
<td>.378</td>
<td>-1.442</td>
<td>-.229</td>
</tr>
<tr>
<td>Post-test</td>
<td>Non-TA</td>
<td>.254</td>
<td>.378</td>
<td>0.672</td>
<td>-.306</td>
</tr>
<tr>
<td></td>
<td>Think-Aloud</td>
<td>.091</td>
<td>.378</td>
<td>0.241</td>
<td>-.790</td>
</tr>
<tr>
<td>Delayed post-test</td>
<td>Non-TA</td>
<td>.604</td>
<td>.378</td>
<td>1.598</td>
<td>-.280</td>
</tr>
<tr>
<td></td>
<td>Think-Aloud</td>
<td>.382</td>
<td>.378</td>
<td>1.011</td>
<td>-.220</td>
</tr>
</tbody>
</table>

<sup>§</sup>Performed with the Lilliefors correction
Table L.3. Tests of homogeneity and sphericity assumptions for the reactivity question

<table>
<thead>
<tr>
<th>Test</th>
<th>F</th>
<th>df1</th>
<th>df2</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>.267</td>
<td>1</td>
<td>76</td>
<td>.607</td>
</tr>
<tr>
<td>Post-test</td>
<td>1.554</td>
<td>1</td>
<td>76</td>
<td>.216</td>
</tr>
<tr>
<td>Delayed post-test</td>
<td>.440</td>
<td>1</td>
<td>76</td>
<td>.509</td>
</tr>
</tbody>
</table>

Mauchly’s Test of Sphericity

<table>
<thead>
<tr>
<th>Within-Subjects Effect</th>
<th>Mauchly’s W</th>
<th>Approx. Chi-Square</th>
<th>df</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>.860</td>
<td>11.330</td>
<td>2</td>
<td>.003</td>
</tr>
</tbody>
</table>

Table L.4. Repeated-measures ANOVA on test scores, comparing the think-aloud conditions

<table>
<thead>
<tr>
<th>Source</th>
<th>Type II Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta²</th>
<th>Obs. Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>3006.588</td>
<td>1.754</td>
<td>1714.063</td>
<td>14.667</td>
<td>.000</td>
<td>.162</td>
<td>.997</td>
</tr>
<tr>
<td>Time x Think-Aloud</td>
<td>426.558</td>
<td>1.754</td>
<td>243.182</td>
<td>2.081</td>
<td>.135</td>
<td>.027</td>
<td>.394</td>
</tr>
<tr>
<td>Error(Time)</td>
<td>15579.344</td>
<td>133.309</td>
<td>116.866</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test of Between-Subjects Effects

<table>
<thead>
<tr>
<th>Source</th>
<th>Type II Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta²</th>
<th>Obs. Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>962641.342</td>
<td>1</td>
<td>962641.342</td>
<td>3511.559</td>
<td>.000</td>
<td>.979</td>
<td>1.000</td>
</tr>
<tr>
<td>Think-Aloud</td>
<td>25.364</td>
<td>1</td>
<td>25.364</td>
<td>.093</td>
<td>.762</td>
<td>.001</td>
<td>.060</td>
</tr>
<tr>
<td>Error</td>
<td>20834.264</td>
<td>76</td>
<td>274.135</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Performed with Greenhouse-Geisser corrections
Appendix M
Latency of thinking aloud

In addition to asking whether thinking aloud has an effect on the content and outcomes of participants’ task performance in terms of accuracy/success, strategies employed, etc., researchers also often check whether it affects the amount of time participants spend. This is referred to as reactivity for time, or latency.

For the latency analyses in the present experiment, 2 participants from each group (i.e., 2 Think-Aloud and 2 Non-Think-Aloud) were excluded. A preliminary inspection of the data had revealed that 2 participants in the Non-TA condition were extreme outliers, having spent more than twice the other participants’ average amount of time in the first two treatment sessions, and continuing to spend almost twice the average amount of time in the third session. The TA condition did not have any similarly extreme outliers, and it was not always the same participants in that condition who tended to spend longer across all treatment sessions than the others; nonetheless, it seemed inequitable to exclude 2 participants from the Non-TA condition while excluding none from the TA condition, particularly since these Non-TA participants’ behavior was contrary to the usual finding that think-aloud conditions tend to take longer. Thus, the decision was made to exclude the 2 participants who took the longest amount of time, on average, in each condition. In the Non-TA condition, these were P54, who spent 101 minutes on average, and P131, who spent 98 minutes on average, both of whom were in the Trees-Feedback condition. In the TA condition, these were P65 (69 minutes, Trees-Feedback condition) and P124 (70 minutes, No-Feedback condition). With the data of these 4 participants excluded, new data points can be identified as outliers (see Figure M.1). However, not being as extreme as those already excluded, and not as substantially affecting the distributional characteristics of the data, they were included in the latency analyses.

Table M.1 and Figure M.1 display descriptive statistics, while Table M.2 presents other information relevant to the normality assumption: skewness and kurtosis ratios and the Kolmogorov-Smirnov and Shapiro-Wilk tests. The goodness-of-fit tests suggest some evidence of non-normal distributions in the Non-TA condition for Treatments 2 and 3. All of the skewness and kurtosis ratios are within the acceptable range of +/-2, however, as can be confirmed by inspecting the boxplots in Figure M.1. None of the deviations look extreme. Table M.3 shows the results of Levene’s tests of homogeneity of variances and Mauchly’s test of sphericity. The fact that Levene’s test produced a statistically significant result for the first treatment means that the variances in the TA and Non-TA groups should not be considered equivalent for that session (see also Table M.1 and Figure M.1, which shows a wider IQR and overall range for the TA condition in Treatment 1). The statistically significant result for Mauchly’s test means that the sphericity assumption was violated; individual participants’ variances cannot be considered equal across the repeated measures. Therefore, the Greenhouse-Geisser correction was used.
**Figure M.1.** Boxplots of amount of time spent per treatment by the TA and Non-TA groups

**Figure M.2.** Average amount of time spent per treatment session by the TA and Non-TA groups
Table M.1. Descriptive statistics of time spent during the treatment sessions by the Think-Aloud and Non-Think-Aloud groups

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Group</th>
<th>Mean</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Std. Dev.</th>
<th>Variance</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n=38 per group)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td>Session 1</td>
<td>Non-Think-Aloud</td>
<td>41.82</td>
<td>49</td>
<td>21</td>
<td>70</td>
<td>11.624</td>
<td>135.127</td>
<td>37.99</td>
</tr>
<tr>
<td></td>
<td>Think-Aloud</td>
<td>51.87</td>
<td>68</td>
<td>22</td>
<td>90</td>
<td>17.783</td>
<td>316.225</td>
<td>46.02</td>
</tr>
<tr>
<td>Session 2</td>
<td>Non-Think-Aloud</td>
<td>40.58</td>
<td>47</td>
<td>23</td>
<td>70</td>
<td>11.663</td>
<td>136.034</td>
<td>36.75</td>
</tr>
<tr>
<td></td>
<td>Think-Aloud</td>
<td>42.84</td>
<td>44</td>
<td>20</td>
<td>64</td>
<td>12.078</td>
<td>145.866</td>
<td>38.87</td>
</tr>
<tr>
<td>Session 3</td>
<td>Non-Think-Aloud</td>
<td>33.29</td>
<td>37</td>
<td>17</td>
<td>54</td>
<td>9.475</td>
<td>89.779</td>
<td>30.18</td>
</tr>
<tr>
<td></td>
<td>Think-Aloud</td>
<td>36.37</td>
<td>48</td>
<td>20</td>
<td>68</td>
<td>10.440</td>
<td>108.996</td>
<td>32.94</td>
</tr>
</tbody>
</table>
Table M.2. Goodness-of-fit tests to check normality of distribution of time spent during the treatments by the TA and Non-TA groups

<table>
<thead>
<tr>
<th>Treatment Group</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Skew</td>
<td>Std. Error</td>
<td>Ratio</td>
<td>Kurtosis</td>
</tr>
<tr>
<td>Session 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-TA</td>
<td>.426</td>
<td>.383</td>
<td>1.112</td>
<td>-.106</td>
</tr>
<tr>
<td>Think-Aloud</td>
<td>.456</td>
<td>.383</td>
<td>1.191</td>
<td>-.589</td>
</tr>
<tr>
<td>Session 2</td>
<td>.717</td>
<td>.383</td>
<td>1.872</td>
<td>-.153</td>
</tr>
<tr>
<td>Non-TA</td>
<td>.039</td>
<td>.383</td>
<td>0.102</td>
<td>-1.057</td>
</tr>
<tr>
<td>Think-Aloud</td>
<td>.539</td>
<td>.383</td>
<td>1.407</td>
<td>.749</td>
</tr>
<tr>
<td>Session 3</td>
<td>.655</td>
<td>.383</td>
<td>1.710</td>
<td>-.190</td>
</tr>
<tr>
<td>Non-TA</td>
<td>.539</td>
<td>.383</td>
<td>1.407</td>
<td>.749</td>
</tr>
</tbody>
</table>

§Performed with the Lilliefors correction
Table M.3. Tests of homogeneity of variances and sphericity assumptions for latency

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Levene’s Test of Homogeneity of Variances</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Levene’s Test of Homogeneity of Variances</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>$F$</td>
<td>df1</td>
</tr>
<tr>
<td>------------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Session 1</td>
<td>7.296</td>
<td>1</td>
</tr>
<tr>
<td>Session 2</td>
<td>.477</td>
<td>1</td>
</tr>
<tr>
<td>Session 3</td>
<td>.673</td>
<td>1</td>
</tr>
</tbody>
</table>

Mauchly’s Test of Sphericity

<table>
<thead>
<tr>
<th>Within-Subjects Effect</th>
<th>Mauchly’s W</th>
<th>Approx. Chi-Square</th>
<th>df</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session</td>
<td>.814</td>
<td>15.049</td>
<td>2</td>
<td>.001</td>
</tr>
</tbody>
</table>

Table M.4. RM ANOVA comparing time spent in treatments by the TA and Non-TA groups

Tests of Within-Subjects Effects

<table>
<thead>
<tr>
<th>Source</th>
<th>Type II Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>$F$</th>
<th>Sig.</th>
<th>Partial Eta$^2$</th>
<th>Obs. Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session</td>
<td>5522.798</td>
<td>1.686</td>
<td>3275.804</td>
<td>67.700</td>
<td>.000</td>
<td>.478</td>
<td>1.000</td>
</tr>
<tr>
<td>Session x Think-Aloud</td>
<td>696.500</td>
<td>1.686</td>
<td>413.123</td>
<td>8.538</td>
<td>.001</td>
<td>.103</td>
<td>.940</td>
</tr>
<tr>
<td>Error(Session)</td>
<td>6036.702</td>
<td>124.759</td>
<td>48.387</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test of Between-Subjects Effects

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>$F$</th>
<th>Sig.</th>
<th>Partial Eta$^2$</th>
<th>Obs. Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>385649.689</td>
<td>1</td>
<td>385649.689</td>
<td>1003.155</td>
<td>.000</td>
<td>.931</td>
<td>1.000</td>
</tr>
<tr>
<td>Think-Aloud</td>
<td>1500.987</td>
<td>1</td>
<td>1500.987</td>
<td>3.904</td>
<td>.052</td>
<td>.050</td>
<td>.496</td>
</tr>
<tr>
<td>Error</td>
<td>28448.325</td>
<td>74</td>
<td>384.437</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Performed with Greenhouse-Geisser corrections
As suggested by Figure M.2, a relatively large and statistically significant difference of 10 minutes between the Think-Aloud and Non-Think-Aloud groups in Session 1 (TA: 52 min, Non-TA: 42 min; \(t(63.739) = -2.917^*, p = 0.005, d = 0.73\)) shrank to a statistically non-significant difference of only 2 minutes by Session 2, apparently due to a tendency on the part of TA participants to complete the activity faster than before (TA: 43 min, Non-TA: 41 min; \(t(74) = -0.831, p = .41, d = 0.19\)). From Session 2 to Session 3, both groups seemed to speed up further to roughly the same degree, resulting in a statistically non-significant difference of about 3 minutes between the groups (TA: 36 min, Non-TA: 33 min; \(t(74) = -1.346, p = .18, d = 0.31\)). In other words, the Think-Aloud group took more time than the Non-TA group to complete the first treatment activity, but not to complete the second or third. All in all, the TA group sped up by about 15.5 minutes from Session 1 to Session 3, while the Non-TA group sped up by about 8.5 minutes. The fastest participants in each group at each session completed the activities at about the same speed: 17-23 minutes per treatment for participants who were not thinking aloud and 20-22 minutes per treatment for participants who were thinking aloud.

In order to determine whether there was a difference between the groups in the amount of time they spent during the treatments overall, and whether they sped up to different extents over the course of the treatments, a RM ANOVA was run with Session (1, 2, 3) as the within-subjects variable and Think-Aloud group (TA or Non-TA) as the between-subjects variable. Those results are presented in Table M.4. The statistically significant effect for Session \((F(1.69, 124.76) = 67.700^*, p < 0.001, \eta_p^2 = 0.48)\) indicates that, considered together, the participants spent fewer minutes per session over time. The between-subjects effect for Think-Aloud group closely approaches statistical significance \((F(1, 74) = 3.904^o, p = 0.052, \eta_p^2 = 0.05)\), but with a weak effect size (only 5% of the variance accounted for) and low power (only a 50% chance of finding a true difference between the groups, if one existed). The statistically significant Session-by-Think-Aloud interaction \((F(1.69, 124.76) = 8.538^*, p = 0.001, \eta_p^2 = 0.10)\) indicates that the TA and Non-TA groups’ trajectories of speeding up across the sessions differed from each other, with the TA group speeding up more dramatically. Running a repeated-measures ANOVA for each group separately, however, the decreases in the amounts of time spent per session over the course of all 3 treatments were statistically significant for learners in both the TA condition \((F(1.42, 52.64) = 42.401, p < 0.001, \eta_p^2 = 0.53)\) and the Non-TA condition \((F(2.00, 73.93) = 29.588, p < 0.001, \eta_p^2 = 0.44)\), with large effect sizes for both.

Interestingly, if the treatment conditions are considered separately, evidence for a latency effect is found in the Trees-Feedback condition, but not in the others. When RM ANOVAs are run for just the No-Feedback condition and just the Right/Wrong-Feedback condition, the effects for Session are statistically significant, indicating that the learners in each of those conditions sped up over the course of the treatments (No-FB: \(F(1.80, 39.66) = 16.047, p < 0.001^*, \eta_p^2 = 0.42; R/W: F(1.67, 45.18) = 21.096, p < 0.001^*, \eta_p^2 = 0.44\)). However, the between-groups TA effects are not statistically significant (No-FB: \(F(1, 22) = 0.361, p = 0.55, \eta_p^2 = 0.02; R/W: F(1, 27) = 0.305, p = 0.59, \eta_p^2 = 0.01\)), nor are the Session-by-TA interaction effects (No-FB: \(F(1.80, 39.66) = 2.320, p = 0.12, \eta_p^2 = 0.10; R/W: F(1.67, 45.18) = 2.694, p = 0.09, \eta_p^2 = 0.09\)). Within the No-Feedback condition and the Right/Wrong-Feedback condition, independent-samples t-tests comparing the Think-Aloud groups for each session separately do not produce statistically significant results, either.
In the Trees-Feedback condition, on the other hand, thinking aloud did affect the amount of time learners spent during the treatments. A repeated-measures ANOVA indicates a statistically significant effect for Session ($F(1.49, 31.29) = 33.622, p < 0.001^*, \eta^2 = 0.62$), indicating that the learners sped up over the course of the treatments; a statistically significant effect for Think-Aloud group ($F(1, 21) = 5.119, p = 0.03^*, \eta^2 = 0.20$), indicating that, overall, the TA and Non-TA groups differed from each other in the amount of time they spent; and a Session-by-Think-Aloud interaction effect that approaches statistical significance ($F(1.49, 31.29) = 3.415, p = 0.058, \eta^2 = 0.14$), suggesting that the TA and Non-TA groups’ trajectories of speeding up over the course of the treatments may have differed. Independent-samples $t$-tests locate differences between the TA and Non-TA conditions for Treatments 1 and 3, but not for Treatment 2 (Treatment 1: $t(21) = -2.487, p = 0.02^*$; Treatment 2: $t(21) = -1.732, p = 0.10$; Treatment 3: $t(21) = -2.094, p = 0.049^*$). Based on these results, it is possible to conclude that when learners received tree diagrams as feedback, verbalizing their thoughts concurrently took more time than completing the activities silently.
Appendix N
Parallel coordinate plots showing test performance of each learner

Treatment condition: No feedback
Appendix O
Boxplots and histograms for individual-difference measures

Appendix O1
Visual Patterns Test

![Boxplot for Visual Patterns Test](image-url)
Appendix O2
Modern Language Aptitude Test

Part IV-Words in Sentences (left) and Part V-Paired Associates (right)
Appendix O3
Test of sensitivity to ambiguity and familiarity with metalinguistic terminology

[Box plot showing distribution of scores by treatment condition]

[Histograms showing frequency distribution of scores]
Appendix O4
Test of sensitivity to ambiguity in English

![Box plots for total score on test of sensitivity to ambiguity in English for different treatment conditions: No feedback, Right-wrong feedback, and Trees feedback.](image)
Appendix O5
Test of familiarity with metalinguistic terminology

![Box plots showing score distribution across different feedback conditions.](image)
Appendix O6
Years of Japanese study

![Boxplot and histograms showing the distribution of years of Japanese study across different feedback conditions.](image)

- No feedback
- Right/wrong feedback
- Trees feedback

![Histograms showing frequency distribution of years of Japanese study.](image)
Linearity of correlations between years of Japanese study and zibun post-test scores

*Note: Outliers are circled.*
Appendix O7
Japanese course levels

No feedback
Right/wrong feedback
Trees feedback

Current level of university Japanese study

Frequency

Current level of university Japanese study
Appendix O8
Numbers of linguistics courses taken

![Graph showing numbers of linguistics courses taken under different feedback conditions. The graph displays data points for 'No feedback', 'Right/wrong feedback', and 'Trees feedback' treatment conditions.]
Appendix O9
Reported enjoyment of grammar

![Box plots and frequency histograms showing enjoyment of grammar across different treatment conditions.](image-url)
Appendix P
Correlations by group between individual-difference measures and zibun test scores

Appendix P1
Test of sensitivity to ambiguity with reflexives
Appendix P2
Test of sensitivity to ambiguity

PRE-TEST

POST-TEST

Score on test of sensitivity to ambiguity in English

Score on pre-test (percentage)

Score on posttest (percentage)

Treatment condition
- No feedback
- Right/Wrong
- Trees feedback

No feedback: R²
Linear = 0.003

Right/Wrong: R²
Linear = 0.172

Trees feedback: R²
Linear = 0.072
Appendix P3
Test of familiarity with metalinguistic terminology

PRE-TEST

POST-TEST

<table>
<thead>
<tr>
<th>Score on pre-test (percentage)</th>
<th>Score on post-test (percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

Treathment condition:
- No feedback
- Right/wrong
- Trees feedback

No feedback: R²
Linear = 0.032
Right/wrong: R²
Linear = 0.008
Trees feedback: R²
Linear = 0.022

No feedback: R²
Linear = 0.178
Right/wrong: R²
Linear = 0.021
Trees feedback: R²
Linear = 0.033
Appendix P4
Test of sensitivity to ambiguity and familiarity with metalinguistic terminology

PRE-TEST

POST-TEST
Appendix P5
Years of Japanese study

*Note: Outliers are circled.
Appendix P6
Japanese course levels

PRE-TEST

POST-TEST

Treatment condition

PRE-TEST

POST-TEST
Appendix P7
Numbers of linguistics courses taken

PRE-TEST

POST-TEST
Appendix P8
Reported enjoyment of grammar

PRE-TEST

POST-TEST

Treatment condition
- No feedback
- Right/wrong
- Trees feedback

No feedback: $R^2$ Linear = 0.1
Right/wrong: $R^2$ Linear = 0.04
Trees feedback: $R^2$ Linear = 0.258

Treatment condition
- No feedback
- Right/wrong
- Trees feedback

No feedback: $R^2$ Linear = 0.005
Right/wrong: $R^2$ Linear = 0.204
REFERENCES


research on language learning and teaching (pp. 133–164). Philadelphia: John Benjamins.


