PUSHING FOR PROCESSING:
The Roles of Depth of Processing, Working Memory, and Reactivity on Comprehension

A Dissertation
submitted to the Faculty of Spanish and Portuguese
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Spanish Linguistics

By
Johnathan D. Mercer, M.S.

Washington, D.C.
December 16, 2014
ABSTRACT
VanPatten’s (1996, 2004) Primacy of Meaning Principle claims that second language learners process for meaning before they process for form. Previous research has empirically tested this principle with varied results (Greenslade, Bowden, & Sanz, 1999; Leow, Hsieh, & Moreno, 2008; Morgan-Short, Heil, Botero-Moriarty, & Ebert, 2012; VanPatten, 1990; Wong, 2001). In each of these studies, attention to form has been operationalized either by circling the specific forms (for the reading modality, e.g., Greenslade et al., 1999; Leow et al., 2008; Wong, 2001) or by placing a check mark on a piece of paper (for the aural modality, e.g., VanPatten, 1990; Wong, 2001). As Leow et al. (2008) note, in their study, this resulted in a low level of processing that may not have been sufficient to have the detrimental effect on meaning postulated in the Primacy of Meaning Principle. Morgan-Short et al. (2012), who conceptually replicated Leow et al. (2008) with the addition of a Non-Think-Aloud group and a larger group of participants, found that level of processing was positively related to comprehension. Nonetheless, as they coded for processing after the experiment was over, their processing conditions were not random, and the potential for mediating variables to have played a role cannot be excluded. One such variable may be working memory, as this variable has been found to be related to reading...
comprehension (e.g., Harrington & Sawyer, 1992), the assessment task, and multitask performance (e.g., König, Buhner, & Murling, 2005). In this study, I randomly assigned participants to six groups, partitioned by Depth of Processing (DP), which included three depths, and the Think-Aloud vs. Non-Think-Aloud groups (TANTA). Data gathered revealed that processing for form at the depth of interpreting negatively affected L2 comprehension. Positive reactivity was also found. No evidence was found to support a role for amount of processing, a relationship between working memory capacity (WMC) and comprehension, or any interaction between the three main variables (DP, WMC, TANTA). The negative impact of processing for form at the depth of interpreting on L2 reading comprehension supports VanPatten’s (2004) Primacy of Meaning Principle.
ACKNOWLEDGEMENTS

Foremost, I would like to express my most gracious thanks to Ron. Your patience, support, and guidance as a professor, director, mentor, and friend were quintessential to the successful completion of my doctoral degree. Thank you for taking the time to share your expertise with me and thoughtfully answer my many, many questions. On a more personal note, your kindness, once even inviting me to your home for Thanksgiving, was heartfelt. I will forever be grateful for all you have done for me. You have my eternal admiration.

I would like to thank Dr. Rusan Chen, my professor and one of my committee members, for both teaching me and guiding me to use the ideal statistical analyses, and Dr. Nina Moreno, also a member of my committee, for her many questions, comments, and corrections. I am most obliged to you both for taking the time to help me mold this dissertation into the product it is today.

Thank you to the many inspirational professors I have had the pleasure to learn from at Georgetown University.

I thank you, Father, very much for your countless hours developing the computer program for the tasks of the second session of the experiment. Without this program, many of the methodological improvements would not have been possible.

Thank you to Dr. Jaemyung Goo for kindly sharing his OSPAN task with me.

I would like to thank Amy Cava and Dr. Sergio Adrada for taking the time to assist me in the laboratory.

I would like to thank Nelia Gustafson and Peter Janssens for their help reserving the laboratory. Thank you also, Peter Janssens and Jonathan Lagdameo, for your technical assistance.

Thank you to Dr. Héctor Campos and Natalia Curto Garcia-Nieto for their help editing my modifications to the experimental texts.

To my participants: You rock!

On a personal level, I thank my Mother and Father for all the sacrifices they made so that I could have the best education possible. I also thank you both for your constant love and support.

I thank my Grandmother for her undying encouragement and love. Your infinite letters, phone calls, and messages have meant so much. You can call me Dr. now 😊

I thank my Grandfather for all the board games we played together, for the leg-wrestling matches, and for the petting zoo trips! You were a great Grandpa! I wish you were here.

Finally, I thank my wife, Liz, mi chiquita y el amor de mi vida, for her patience, understanding, and love. Every day, no matter how much work I had to do, being able to count on you made it special. I thank you for the delicious Peruvian-style home-cooked meals you made me almost every day, for your sacrifice of countless weekends, and for the hugs and kisses you showered me with as soon as I opened the door each day I came home. Te amo.
In dedication to my Father and to my Wife.
# PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>INTRODUCTION</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement of the Problem</td>
<td>1</td>
</tr>
<tr>
<td>Definition of Terms</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 1: ATTENTION</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>8</td>
</tr>
<tr>
<td>Attention as Limited</td>
<td>9</td>
</tr>
<tr>
<td>Cognitive Models of Attention</td>
<td>9</td>
</tr>
<tr>
<td>Models of Attention in SLA</td>
<td>28</td>
</tr>
<tr>
<td>Second Language as a Cognitive Skill</td>
<td>28</td>
</tr>
<tr>
<td>Attention as Alertness, Orientation, and Detection</td>
<td>30</td>
</tr>
<tr>
<td>The Noticing Hypothesis</td>
<td>33</td>
</tr>
<tr>
<td>Empirical Evidence</td>
<td>33</td>
</tr>
<tr>
<td>Attention and Memory</td>
<td>49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 2: INPUT PROCESSING</th>
<th>51</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>51</td>
</tr>
<tr>
<td>Models of processing in SLA</td>
<td>52</td>
</tr>
<tr>
<td>VanPatten's Primacy of Meaning Principle</td>
<td>57</td>
</tr>
<tr>
<td>Empirical Research</td>
<td>58</td>
</tr>
<tr>
<td>Depth of Processing</td>
<td>67</td>
</tr>
<tr>
<td>Empirical Evidence &amp; Depth of Processing in the L1</td>
<td>71</td>
</tr>
<tr>
<td>Methodological Issues of Empirical Research</td>
<td>98</td>
</tr>
<tr>
<td>Working Memory</td>
<td>102</td>
</tr>
<tr>
<td>Working Memory in Cognitive Science</td>
<td>103</td>
</tr>
<tr>
<td>Measurements of Working Memory Capacity</td>
<td>105</td>
</tr>
<tr>
<td>Models of Working Memory</td>
<td>107</td>
</tr>
<tr>
<td>Working Memory and Language</td>
<td>114</td>
</tr>
<tr>
<td>Empirical Research</td>
<td>115</td>
</tr>
<tr>
<td>Working Memory and L1 Reading Comprehension</td>
<td>119</td>
</tr>
<tr>
<td>Working Memory and L2 Reading Comprehension</td>
<td>121</td>
</tr>
<tr>
<td>Working Memory and Multitask Performance</td>
<td>125</td>
</tr>
<tr>
<td>Concurrent Verbal Reports</td>
<td>130</td>
</tr>
<tr>
<td>Concurrent Verbal Reports in SLA</td>
<td>132</td>
</tr>
<tr>
<td>Think-Alouds: Measuring Depth of Processing</td>
<td>136</td>
</tr>
<tr>
<td>Reactivity</td>
<td>139</td>
</tr>
<tr>
<td>Ericsson and Simon's (1996) Theoretical Analysis on the Validity of Think-Alouds</td>
<td>139</td>
</tr>
<tr>
<td>Empirical Studies in SLA</td>
<td>140</td>
</tr>
<tr>
<td>Reactivity and Working Memory Capacity</td>
<td>145</td>
</tr>
<tr>
<td>Reactivity and Metacognitive Concurrent Verbal Reports</td>
<td>147</td>
</tr>
<tr>
<td>Purpose of the Study</td>
<td>149</td>
</tr>
<tr>
<td>Research Questions</td>
<td>152</td>
</tr>
</tbody>
</table>

| CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY | 153 |

vii
# Introduction to the Present Study

**Participants**

**Linguistic Target**

**Materials**

**Assessment Materials**

**Procedure**

**Elimination Criteria**

**Scoring**

**Quantitative Analysis**

**Qualitative Analysis**

**Discussion**

**Minor Changes**

**CHAPTER 4: RESULTS**

**Qualitative Analysis**

**Processing For Meaning**

**Skipping**

**Time Eliminations**

**Participant Strategies and Commentary**

**Quantitative Analysis**

**Preliminary Analysis**

**Research Questions I, II, III**

**Research Question IV**

**Research Questions V, VI, VII**

**CHAPTER 5: DISCUSSION AND CONCLUSION**

**Introduction**

**Discussion**

**Research Question I**

**Research Question II**

**Research Question III**

**Research Question IV**

**Research Question V**

**Research Question VI**

**Research Question VII**

**Pedagogical Implications**

**Future Research**

**Conclusion**
LIST OF FIGURES

Figure 1. Descriptive Statistics 200
Figure 2. Depth of Processing and TANTA, Relationship to Multiple-Choice Comprehension Score 203
LIST OF TABLES

Table 1. Participants by Condition and TANTA .................................................. 159
Table 2. Comprehension Score by Spanish Class Level ....................................... 164
Table 3. Overall Descriptive Statistics .................................................................. 165
Table 4. Overall Descriptive Statistics .................................................................. 166
Table 5. Participants by Condition and TANTA .................................................. 184
Table 6. Comprehension Score by Spanish Class Level ....................................... 198
Table 7. Impact of Depth of Processing and TANTA on Comprehension ............... 199
Table 8. Overall Descriptive Statistics .................................................................. 200
Table 9. Post-Hoc Scheffé Analysis ...................................................................... 201
Table 10. Coefficient Variables of Multiple Regression Analysis ......................... 203
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

INTRODUCTION

Statement of the Problem

VanPatten (1996, 2004) theoretically explores the limits to human processing, which he defines as making a connection between meaning and form. VanPatten’s (2004) first principle, the Primacy of Meaning Principle, holds that input is first processed for meaning, and second for form. As processing resources are assumed to be constrained by working memory, the Primacy of Meaning Principle implies that processing for form may be limited in situations where learners expend too many of their processing resources for meaning in processing lexical items and content words. The principle also implies that when learners are given tasks that require them to expend high levels of processing resources to process for form and meaning at the same time, without allowing them any additional time to compensate for the added burden of processing for form, that their comprehension will be negatively affected. Previous research has empirically tested this implication with varied results (Greenslade, Bouden, & Sanz, 1999; Leow, Hsieh, & Moreno, 2008; Morgan-Short, Heil, Botero-Moriarty, & Ebert, 2012; VanPatten, 1990; Wong, 2001).

Although VanPatten (1990) and Greenslade et al.’s (1999) results provided support for the principle, Wong (2001) only found attention to the *article* to have a significant impact on comprehension, and only for listening. More recently, Leow et al. (2008), who employed concurrent verbal reports to establish that participants were processing for meaning, did not find any of their form conditions to have a significant impact on comprehension. Morgan-Short et al. (2012), however, did find a positive relationship between processing level and comprehension.
Nonetheless, when interpreting the results of Leow et al. (2008) and Morgan-Short et al. (2012), important methodological concerns must be considered.

First, in Leow et al. (2008), only written instructions were used to ensure that participants completed the tasks as quickly as possible, and in Morgan-Short et al. (2012) no attempt was made to ensure that participants did so. One crucial aspect of VanPatten’s input processing model, however, is that learners will have difficulty processing for form and meaning simultaneously when cognitive resources are limited. With regard to Morgan-Short et al. (2012), who did not pressure participants to complete the reading quickly, participants in the greater depth of processing conditions may have taken additional time to access the cognitive resources necessary to compensate for the potentially burdensome task of processing for form. Even in Leow et al. (2008), who instructed participants to read as quickly as possible, said instructions may have been insufficient for participants to indeed complete the tasks as speedily as they were able.

Second, Leow et al. (2008) and Morgan-Short et al.’s (2012) exploration of levels of processing may have been methodologically problematic. Leow et al. (2008) and Morgan-Short et al. (2012) explored the potential impact of three different levels of processing that were coded based on participants’ concurrent verbal reports. The three levels were: (I) Only circling; (II) Circling and reporting; (III) Circling, reporting and interpreting. In Leow et al. (2008), while the researchers found no evidence of a link between higher levels of processing and comprehension score, they caution that very few participants were in each cell. Morgan-Short et al., who conceptually replicated Leow et al. with the addition of a Non-Think-Aloud group and a larger base of participants, found that level of processing was positively related to comprehension. Leow et al. (2008) and Morgan-Short et al.’s (2012) coding scheme for processing levels,
however, may have been flawed. In Morgan-Short et al. (2012), participants only needed to process one item at a greater level to be coded as having processed at that level, while Leow et al. (2008) coded participants as processing at level III based on their translation of “very few items” out of a total of 10 (p. 681). Moreover, at least one of the criteria that Leow et al. (2008) and Morgan-Short et al. (2012) used to separate level I and level II participants, whether participants circled the forms or they circled and verbalized them, may not have been sufficient to ensure a difference in processing. Finally, given that levels of processing were coded after the experiment was over based on the think-aloud verbal reports in both studies, their processing groups were not random, and the potential for mediating variables to have played a role cannot be excluded.

One mediating variable may be working memory, a variable that has been found to be related to both reading comprehension (e.g., Daneman & Carpenter, 1980; Harrington & Sawyer, 1992) and multitask performance (e.g., Colom, Martínez-Molina, Shih & Santacreu, 2010; König, Bühner, & Mürling, 2005). Given that previous studies on processing for form and meaning addressed reading comprehension, and involved a “dual task” (Morgan-Short et al., 2012, p. 666), in that participants needed to process both meaning and form at the same time, working memory could potentially have played a role in the incongruosity of previous findings.

A methodological issue specific to Morgan-Short et al. (2012) was that the Non-Think-Aloud control group may have been compromised. Morgan-Short et al. found the Think-Aloud condition to have a significant impact on the results, although they emphasize that the effect size was minimal ($r^2 = 0.01$) to be of practical value, citing Ferguson (2009, p. 533). This finding, however, needs further validation, as Morgan-Short et al. eliminated those participants who
backtracked (reread the text) in the Think-Aloud condition, but not in the Non-Think-Aloud condition, potentially reducing the comparative power between the two groups.

In this study, I revisited the empirical investigation of the Primacy of Meaning Principle by operationalizing three depth of processing levels that were randomly assigned to participants, as opposed to subsequently coded based on Think-Aloud data. The first depth of processing was processing for meaning only. The second level of processing was at the depth of identifying (clicking on past forms), which is based on Leow et al.’s (2008) first and second levels of processing (see above). The third level of processing, at the depth of interpreting (labeling past forms as preterit or imperfect), was based on Leow et al.’s (2008) third level of processing (circling, reporting, and interpreting). In addition, I employed a time-pressure countdown clock, followed by a countdown time-limit, in order to limit participants’ ability across depth of processing levels to access additional cognitive resources. Finally, I examined the potential roles of reactivity and working memory capacity in processing for meaning and form simultaneously.
Definition of Terms

Attention: The readiness and orientation of cognitive resources to the input (Tomlin & Villa, 1994). Cognitive and SLA researchers generally agree that attention is capacity limited, although disagreement exists as to the nature of this limit (Broadbent, 1958; Tomlin & Villa, 1994; Treisman, 1960; Robinson, 1995).

Awareness: A cognizance of an aspect of the input to which the learner is exposed. According to Tomlin and Villa (1994), awareness “refers to a particular state of mind in which an individual has undergone a specific subjective experience of some cognitive content or external stimulus” (p. 193). Schmidt (1990) differentiates between three levels of awareness: (1) Perception; (2) Noticing; and (3) Understanding (p. 132).

Cognitive Resources: Mental capital that an individual has available to complete a task or set of tasks. Cognitive resources are limited (e.g., Broadbent, 1958), and, for any one task or set of tasks, vary as a function of time available for completion and working memory capacity.

Input: Input refers to the linguistic data, written or spoken, to which “L2 learners are exposed” (VanPatten & Benati, 2010, p. 94). Exposure to input does not necessarily imply learning or comprehension.

Intake: Intake refers to the input that a language learner has attended to, detected, and noticed. Intake may be further processed in working memory or may be discarded (Leow, 2015).
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

Learning: Learning is a general term that can refer to intake, production, and/or integration into the L2 system.

Reactivity: Reactivity refers to the potential impact of verbal protocols on performance, that is, “whether the act of thinking aloud alters the end state of the cognitive process (the accuracy of task performance)” (Leow & Bowles, 2005, p. 184). Reactivity may be negative or positive, meaning, respectively, that it may be detrimental to or improve performance.

Processing: Processing refers to the use of cognitive resources to take in input, and/or integrate input into the developing system (Gass & Selinker, 2008).

Processing at the depth of identifying: The use of cognitive resources to find a form or set of forms in the input.

Processing at the depth of interpreting: The use of cognitive resources to interpret, that is, make a mental note of, a form’s meaning and/or grammatical function.

Think-Aloud Protocols: The recorded research data that is obtained as a result of participants thinking aloud their thoughts while they complete a task. Think-Aloud protocols are useful in that they offer the researcher insight into how participants process during a particular task (Leow & Bowles, 2005).
Working Memory Capacity: The degree to which any individual participant is able to simultaneously “maintain and process goal relevant information” (Conway et al., 2007, p. 3). Working memory capacity is an important individual difference variable in SLA, having been found to be related to performance in a number of language learning categories (e.g., Leeser 2007; Walter, 2004).
CHAPTER 1: ATTENTION

Introduction

The nature of attention is key to the understanding of the second language learner as it controls access to linguistic information, intake, and, ultimately, learning (Schmidt, 2001). It is generally agreed upon by cognitive psychologists and second language acquisition researchers that attention is limited (Broadbent, 1958; McLaughlin, 1987; Robinson, 1995; Schmidt, 2001; Treisman, 1960; Kahneman, 1973; VanPatten, 2004). Nevertheless, while Broadbent (1958) argues that attention serves as a one-channel filter through which information must pass if it is to be perceived, Treisman (1960, 1964) argues that attention is not an absolute barrier, and, that under certain circumstances, such as a favorable context, information that is not attended to can indeed be perceived. Empirical evidence suggests, however, that attention is indeed necessary for perception to occur (Broadbent, 1954; Moray, 1959; Treisman, 1960, 1964), although Treisman (1964) disagrees with this interpretation of the evidence.

Attentional models in second language acquisition have another point of contention: whether or not awareness is necessary for intake to occur. Tomlin and Villa’s (1994) model divides attention into three levels (alertness, orientation, and detection), and theorizes that detection, which occurs without awareness, can lead to intake. Schmidt (2001) and Robinson (1995), contrary to Tomlin and Villa (1994), argue that both attention and awareness are necessary for intake. While Leow (1998) provided early evidence in favor of Tomlin & Villa’s (1994) model, Simard & Wong (2001) pointed out important methodological concerns with Leow’s (1998) study, and a critical examination of subsequent empirical evidence (Hama & Leow, 2010; Leow, 1997, 2000; Leung & Williams, 2011; Rosa & Leow, 2004; Rosa & O’Neil,
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

1999; Schmidt & Frota, 1986; Williams, 2005) suggests that awareness may play a role in intake and/or production.

The limited nature of attention, as well as attention being a requirement for perception to occur, provide the necessary theoretical foundation for the current dissertation, which sought to explore the workings of attentional limitations in the context of reading comprehension.

Attention as Limited

Attention is theorized as being of a limited-capacity in both cognitive psychology (e.g., Broadbent, 1958; Treisman, 1960; Kahneman, 1973) and second language acquisition (e.g., McLaughlin, 1987; Robinson, 1995; Schmidt, 2001; VanPatten, 2004). The exact nature of this limit, however, is controversial.

Cognitive Models of Attention

In cognitive science, Broadbent (1958, p. 297) argues that only one “communication channel” can be attended to at any one time, and that input that is not attended to does not pass through the filter, and, thus, will not be perceived. According to Broadbent’s (1958, p. 297) theory, the organism performs a “selective operation” on input available to it. This selective operation determines what input will pass through the filter. Whether a particular input is selected or not, according to the theory, is determined by both aspects of the available input and the “drives” of the organism (p. 298). Input that is salient, such as a highlighted word in a passage, has a greater chance of being attended to, as does input that has been attended to recently. In addition, input that is related to the focus of the organism, such as the main idea of a
passage for an individual reading a newspaper, also has a greater possibility of passing through the attentional filter. Broadbent (1958) argues that once input is selected, a minimum amount of time needs to pass in order for the organism to take action based on the input.

In her attenuated filter model theory, Treisman (1960, 1964), in contrast to Broadbent’s attentional theory, argues that unattended input can be perceived under certain circumstances. According to Treisman’s theory, information that is not attended is not blocked, but, rather, attenuated. In this way, it is possible, but, importantly, more difficult for non-attended information to pass through the filter. For this reason, in order to be analyzed, non-attended information should carry more importance, such as by being an important phrase such as “fire,” or one’s name, or by being contextually important (p. 246). Treisman (1960, 1964) points out empirical evidence, which, according to her, demonstrate instances in which “the barrier of attention” has been bypassed (Moray, 1959; Treisman, 1960, 1964).

Moray (1959) explored the nature of attention in three dichotic listening experiments. In all three experiments, participants were presented with two audio outputs, one in each ear. In the first experiment, a prose message and a short list of words were the two audio outputs, with participants being required to shadow the prose message. A central assumption of the experiment, as well as those to follow (Treisman, 1960, 1964), is that participants pay attention only to the message that they shadow, not to the message in the non-shadowed or ‘rejected’ ear. Participants were given a recognition test to see what words they remembered from the short list of words. On average, participants claimed to recognize more words that had not been previously presented to them than those words that they were actually played to them in the non-shadowed ear. This suggests that any words played to the rejected ear that the participants indicated as
recognizing may have been a result of chance, as opposed to actual recognition. Therefore, these results support Broadbent’s (1958) theory that non-attended information is not perceived.

In experiment 2, Moray (1959) tested participants’ ability to hear instructions from the rejected ear. In the experiment, twelve participants listened to and shadowed ten passages of light fiction in one ear, which were interpolated with the researcher’s instructions in the rejected ear. In half of these instances of instructions, the researcher called the participant by his name. Moray (1959) found that participants either followed or reported hearing the instructions at a significantly higher rate when their name was used (51%), compared to when their name was not mentioned (11%). This result appears to contradict Broadbent’s (1958) theory, as Treisman (1964) suggests, given that if all information that is not attended is not perceived, one would not expect participants to react differently to different types of non-attended information. Nonetheless, Broadbent (1958) accounts for “multi-channel listening” by suggesting that listeners can “change channels” to direct their attention to the “relevant” information (p. 211). In other words, an alternative explanation may be that participants’ names were salient enough to cause participants to momentarily “shift their attention” (Broadbent, 1958, p. 210) to the information from the rejected ear.

Experiment 3 explored the ability of participants to recall digits in the rejected ear. In that experiment, 28 participants were presented with two dichotic messages and instructed to listen to and shadow one of them. The researchers interpolated these messages with numerical digits, and told half the participants that they should try to remember them. Participants then were prompted to recall these digits. In her analysis, Moray (1959) did not find evidence that the numbers were able to “break through the attentional barrier” (p. 59). In Moray’s (1959) three experiments, therefore, only the results of experiment 2 could potentially be interpreted as supporting
Treisman’s (1964) view that, under certain circumstances, information may be perceived without attention.

Treisman (1960) also employed a dichotic listening task in order to test Broadbent’s theory that information must be passed through an attentional filter in order to be perceived. In the experiment, 18 undergraduates put on two headphones, one on each ear, and were asked to shadow what they heard out of one of the two headphones. Four types of passages were played to participants: “(a) narrative passages from a novel; (b) extracts from a technical discussion; (c) eighth order statistical approximations to English; [and] (d) second order approximations to English” (p. 243). The playing of each passage switched ears at approximately the middle of the passage, a point the author refers to as “the break” (p. 244). While only two participants noted that they thought they might have said one or two words from the wrong ear, Treisman (1960) found that 15 of the 18 participants repeated some words from the rejected ear. After the break, these instances were found to be significantly more common for the narrative passages. In her discussion of the findings, Treisman (1960) notes that because participants did not report being aware that they were repeating information from the wrong ear, that it is “unlikely” that they attended to this information (p. 246). Based on this assumption, she then claims that because participants did nonetheless repeat information from the rejected ear that they were indeed able to perceive this “unattended” information, an assertion that is contrary to Broadbent’s (1956) theory that only attended information can be perceived.

Nonetheless, Treisman’s (1960) reasoning for her assumption that participants did not attend to the information they repeated is not convincing. First, the mere fact that participants did not report being aware that they were repeating information from the wrong ear does not necessarily mean that they were not. Given that they listened to and repeated 20 different
passages, it is possible that they may have forgotten that they repeated from the wrong ear by the end of the experiment. Second, even if participants were indeed not aware that they were repeating information from the wrong ear, this only indicates that they were not attending to the location of the information they were repeating, not necessarily the information itself. Indeed, given that these instances were significantly more common after breaks for narrative novel passages, it seems likely that participants simply mistakenly continued to attend to the same passage instead of to the same ear. Broadbent’s (1958) assertion that only attended information can be perceived, therefore, does not appear to be disproven by Moray (1959) or Treisman (1960).

The main components of Broadbent’s Attentional Filter Model (1958), namely that attention is both limited and necessary for learning, have been generally adopted in second language acquisition literature (e.g., McLaughlin, 1987; Schmidt, 1990; VanPatten, 1996, 2004). VanPatten’s (1996, 2004) Input Processing Principles, which serve as the foundation of the current thesis, are one example. The principles postulate that learners prioritize their input processing, such that certain aspects of input, for a variety of reasons, are processed before others. This implies that there is indeed a limit to attention, as otherwise learners would not need to prioritize their processing, as they could simply process for everything. More recent theories of attention have built on and added to Broadbent’s (1958) general model (Carbonell, 1966; Kahneman, 1973; Senders, 1964; Wickens, 2007).

Senders (1964) noted that previous models of attention did not incorporate the impact of expectations on the orientation of attention. That is to say, individuals may orient their attention based on previous experience perceiving the information they desired in a particular sample. To address the potential impact of expectations, Senders (1964) conducted an experiment in which
he observed the eye-movements of five participants. The five participants sat in front of a dialer and were required to press on a trigger whenever any one dialer exceeded a particular established threshold. They were given an incentive to press the trigger within one second after the threshold was exceeded. Senders (1964) found that the eye movements of participants were not random, but, rather, responded to the probabilities that the required information would appear in a particular sample. In this way, Senders (1964) showed that the orientation of attention responds to participants’ expectations.

Building on Senders (1964), Carbonell (1966) argued that, in addition to expectations, people likely orient attention in response to the cost of not paying attention to information in a particular sample. Carbonell (1966), showed, in his computer pilot simulation, that the computer selects the item with the highest cost, makes adjustments to lower the cost, and proceeds to pay attention to the next item. Carbonell, Ward, and Senders (1968) tested Carbonell’s (1966) model by comparing the eye fixations of three experienced Air Force Reserve pilots with the predictions of the model. Carbonell et al. (1968) found a high correlation between the model’s predictions and the eye fixations of the pilots, providing evidence in favor of Carbonell’s (1966) assertion that cost plays a role in the orientation of attention.

Kahenman (1973) posits a role for effort. According to Kahneman’s theory on attention and effort, effort depends primarily upon task demands and time pressure, and, to a lesser extent, on an individual’s intentions. Kahneman (1973) argues that relatively easy tasks are unlikely to involve as much effort as more difficult tasks, even if one is highly motivated to do well on a particular easy task (i.e., has high intentions). Kahneman (1973) offers the following example:
First, try to mentally multiply 83 by 27. Having completed this task, imagine that you are going to be given four numbers, and that your life depends on your ability to retain them for ten seconds. The numbers are seven, two, five and nine (p. 14). Here, Kahneman maintains, it is clear that the first task, which is more demanding, results in more effort, demonstrating the importance of task demands on effort. Kahneman argues that effort determines the portion of total attentional capacity that will be dedicated to a particular activity, and, in turn, performance. Empirical studies support the tenets of Kahneman’s (1973) theory, as they have shown that incentives increase performance (Bahrick, Fitts, & Rankin, 1952) and pupil dilation (Kahneman, Peavler & Onuska, 1968), but that this impact depends on the difficulty of the task (Kahneman et al., 1968).

Aspects of many of the previously discussed models (Broadbent, 1958; Carbonell, 1966; Kahneman, 1973; Senders, 1964) are combined in the Wicken’s SEEV model (Horrey, Wickens, & Consalus, 2006; Wickens, 2007). According to the SEEV model, four factors determine the allocation of attention: (1) Saliency; (2) Effort; (3) Expectancy; (4) Value.

According to the model (Horrey et al., 2006; Wickens, 2007), saliency includes events that stand out, such as “bright lights” or loud “sounds” (Wickens, 2007, p. 179). Broadbent (1958) made a similar claim, theorizing that “physical intensity” had an influence on what events are selected to pass through the attentional channel (p. 297). In support of the importance of salience, Horrey et al. (2006) cite Itti and Koch (2000), a cognitive study in which participants located the target object more quickly when it was more visually salient in terms of color and intensity. Wickens (2007) argues strongly for a role of salience in second language acquisition, citing researchers who claim that learners do not pick up on non-salient features of the input (Ellis, 2006; Robinson, 1995; Schmidt, 2001). While second language acquisition research does
indeed suggest that this may be the case for oral communication (Goldschneider & DeKeyser, 2001), its role in written communication is less clear, as enhanced written input has not generally been found to improve intake (Leow, 1997; Leow, 2001; Leow, Egi, Nuevo & Tsai, 2003; Overstreet, 1998) when compared to unenhanced input.

Goldschneider and DeKeyser (2001) performed a meta-analysis on the importance of perceptual salience, semantic complexity, morphophonological regularity, syntactic category, and frequency in L2 morpheme acquisition. They argue that each of these factors can be considered an aspect of salience. The researchers limited their analysis to studies that examined English as a second language, involved oral data, evaluated functors based on suppliance in obligatory context (SOC), provided sufficient quantitative data, and examined six specific functors: “present progressive –ing; plural –s; possessive ‘s; articles; 3rd person present –s; regular past –ed” (p. 19). In addition, the researchers combined studies on both children and adults, and studies with different types of instruction or exposure to English. In the multiple regression analysis of the data, they found each of the five aspects of saliency contributed significantly to the variance in accuracy of the grammatical functors, suggesting that salience may play a role in L2 attention to oral input. The importance of salience to attention in the written modality has been explored in a number of studies on enhanced input (e.g., Jourdenais, Ota, Stauffer, Boyson, & Doughty, 1995; Leow, 1997; Leow, 2001; Leow, Egi, Nuevo, & Tsai, 2003; Shook, 1994; Overstreet, 1998).

As Leow (2009) points out, the majority of studies on written enhancement that have compared an enhancement group to a non-enhancement group while maintaining other variables constant, what Leow (2009) refers to as “Non-conflated Input Enhancement Studies” (p. 22), have not found enhancement to have a significant impact on intake of target grammatical items
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING (e.g., Leow, 1997; Leow, 2001; Leow et al., 2003; Overstreet, 1998) when compared to the non-enhancement group. Two studies that Leow (2009) classified as “Non-conflated” (Jourdenais, Ota, Stauffer, Boyson, & Doughty, 1995; Shook, 1994) did find input enhancement to have a significant impact on intake when compared to the non-enhancement group. Nonetheless, upon closer inspection, these studies also appear to be conflated. Jourdenais et al. (1995) included participants with a range of scores between 2 and 10 out of a total of 10 possible points, with an average score of 8.14. While the researchers did attempt to balance the enhancement and non-enhancement groups by pre-test score, the significant finding for enhancement cannot be teased apart from the potential impact of prior knowledge. In other words, given the high average pre-test score, it may be that enhancement had a significant impact only on participants with prior knowledge, an argument that Leow (2009) made with regards to Lee (2007). The same argument can be made with regards to the significant finding for enhancement in Shook (1994), as participants who indicated previous study of the grammatical items being examined were nonetheless included in the subject pool. Overall, therefore, the evidence of empirical studies to date on written enhancement strongly suggests that enhancement does not have an impact on participant intake when compared to unenhanced input. In addition, in Leow et al. (2003), no significant difference was found between the enhanced and the unenhanced group for the noticing of target forms. Therefore, while Wickens’ (2007) factor of salience appears to play a role in the allocation of attention in the oral modality of language acquisition (Goldschneider & DeKeyser, 2001), the same may not be the case in the written modality (e.g., Leow et al., 2003; Leow, 2009).

Wickens (2007) describes effort, the first E in the SEEV model of attention, as a movement to shift attention that, when repeated, can result in fatigue. Activities that require
effort are avoided, particularly when the individual is “effort-depleted” (p. 179). Wickens (2007) notes that studies on the operation of motorized vehicles in which effortful activities are avoided by the operator. Recartes and Nunes (2000), for example, found that drivers glanced less frequently at the speedometer and mirrors while completing a concurrent spatial-imagery task. With regards to second language acquisition, Wickens (2007) theorized that effort might play a role in look-up behavior, in which students able to click on words might be more likely to look them up than those with a physical dictionary. Evidence from studies on look-up behavior in second language acquisition does seem to suggest students may opt for the option requiring the least effort, as studies have also found that L2 students prefer looking up L2 definitions or accessing L1 translations to other, potentially more effortful methods to determine the meaning of a word, such as listening to pronunciations (Davis & Lyman-Hager, 1997; Lomicka, 1998), accessing images, references, and/or questions (Lomicka, 1998), or accessing grammar, cultural background information, pictures, and/or a tree denoting antecedent relationships (Davis & Lyman-Hager, 1997). While it is possible that students may have chosen the traditional methods for reasons other than the amount of effort required, such as because they perceived the traditional methods as more accurate, it has also been found that students prefer to look-up only traditional definitions or only translations to looking up both (Davis & Lyman-Hager, 1997; Laufer & Kimmel, 1997; Laufer & Hill, 2000), the latter of which cannot be reasonably perceived as providing less information or being less accurate. The results of studies on look-up behavior, therefore, as Wickens (2007) theorized, do indeed appear to suggest that effort plays a role in the allocation of attention in second language learners.

Expectancy, the second E in the SEEV attentional model, is the tendency to sample the channel that offers the highest bandwidth of relevant information (Horrey et al., 2006). Horrey et
al. (2006) offers the example of a driver, who, in order to maintain a good lane positioning, will tend to visually sample the road information that allows her to compare her lane position with the ideal lane position. A relevant study to the role of expectancy in driving is Tsimhoni and Green (2001), who examined the performance and glances of 16 participants. The participants performed three different map-reading tasks as they drove under five different conditions, which included “parked, straight road, easy curve, moderate curve, and sharp curve” (p. 1586).

Tsimhoni and Green (2001) found that, for longer tasks, the duration of glances to the map was significantly lower than when the curve was sharper. As sharper roads presumably increase the relevance and bandwidth of the road information to a driver seeking to maintain an ideal lane position, the decrease in glance duration to the map information in this situation suggests that expectancy may be playing a role in the increased allocation of attention in driving. The inverse was found in Horrey et al. (2006), where more complex in-video technology (IVT) tasks led participants to sample road information less, providing more evidence of the role of expectancy in driving. Wickens (2007) has difficulty, however, connecting the role of expectancy to second language acquisition, noting that he does not find it more likely that second language learners will pay attention to faster, higher bandwidth, speakers than slower, lower bandwidth, speakers. Perhaps not, but this does not appear to be a fair analogy since, as Wickens (2007) notes, the second language learner may have difficulty understanding the faster speaker. One might argue here, however, that salience and effort may be playing a role, as the understandable data may be more salient and require less effort of the second language learner. Aspects other than those mentioned by Wickens (2007), however, seem to provide a clearer example of the role of expectancy in second language acquisition.
The importance of expectancy to second language learners can be seen in activities that involve multitasking. In a discussion of a reading in the second language, for example, students must balance their attention between listening to the instructor and their fellow students, taking notes, and glancing at the reading. In this particular situation, students may indeed allocate their attention more to their instructor if he/she is covering information that is more relevant or is speaking more quickly. In the same way, students that read a text and are required to pay attention to both the meaning and certain grammatical forms in the text, such as in VanPatten (1990), may vary their allocation of attention according to the bandwidth of the information. If a particular paragraph has numerous grammatical forms, for example, a student may focus more on the grammatical forms than the meaning of the text. Likewise, if a section of the text that is highly concentrated in meaning, such as one that includes a plethora of dates that the student perceives as being crucial to the subsequent comprehension text, a student might allocate more attention to the content aspect of the reading and less to the grammatical forms. To the best of my knowledge, however, the role of expectancy has yet to be empirically examined in the second language literature.

Value, the V in the SEEV model, is the relative worth of samples of information. According to Wickens (2007), people tend to attend to information that has greater value to them given their situation, concerns, and goals. For example, in Horrey et al. (2006), participants allocated significantly more attention to the road, which has a high value in that it helps avoid accidents, than to the in-video technology information, which has a lower value in that it simply helps avoid going in the wrong direction. As Wickens (2007) points out, value is of particular importance to second language acquisition.
VanPatten’s (1996, 2004) processing principals are of particular relevance to this factor in terms of its applicability to second language acquisition. The first of these principles, the Primacy of Meaning Principle, stipulates that learners tend to pay more attention to information that maximizes their access to the meaning of a text, such as content words and lexical items, before other items that are less relevant to meaning, such as redundant words and nonmeaningful grammatical forms (e.g., gender). Given that the assumed goal of the learner is to understand the text, the content words can be seen as having greater value to the learner and, therefore, VanPatten’s Primacy of Meaning Principle is consistent with the value factor of the SEEV model, and reveals the relevance of the factor to second language acquisition. Studies empirically examining the principle, however, have varied in their results (e.g., Greenslade, Bouden, & Sanz, 1999; Leow, Hsieh, & Moreno, 2008; Morgan-Short, Heil, Botero-Moriarty, & Ebert, 2012; Wong, 2001; VanPatten, 1990).

According to the SEEV model, Saliency, Effort, Expectancy, and Value predict the chances of a particular source being sampled. People are most likely to sample sources that are high in saliency, expectancy, and value, and require little effort. In many situations, however, attention to a particular source may be favored by some factors, yet disfavored by others. With regards to the oral modality, a particular speaker of a foreign language may be speaking very quickly about an interesting topic to an audience, thus making this source high in expectancy and value, but the effort required to understand might be so high for some members of the audience, due to the speech rate, that these members may not allocate sufficient attention to the speaker. On the other hand, a speaker that is talking about an uninteresting topic, making the source of little value, but who is speaking slowly, decreasing the effort required to understand, might be attended to more. With regards to the written modality, a second language learner might have a
particular purpose in mind when reading reviews, such as recommending a movie to a friend who loves John Goodman, that he may allocate less attention to paragraphs with a high concentration of details (high expectancy), or are highlighted (high saliency) in order to focus on those paragraphs that specifically talk about John Goodman (high value) and are easier to understand (low effort). According to the theory, therefore, as Wickens (2007) asserts, it is the “weighted sum of the four components” (p. 182) that determines the likelihood that a particular source will be sampled, as second language learners (as well as people in other situations) are forced to sacrifice sources that are favorable according to some factors, in order to sample sources that are more favorable according to the remaining factors. Wickens’ (2007) Multiple Resource Model addresses the ability of individuals to perform tasks that require attention to multiple sources.

According to the Multiple Resource Model (Wickens, 2007), the ability of participants to perform two tasks at the same time depends on the degree to which these two tasks interfere with each other. Wickens (2007) identifies three factors which influence how much one or the other task is negatively affected as a result of multitasking: (1) “effort or resource demands”; (2) “qualitative resource similarity”; and (3) “allocation policy between the two tasks” (p. 183).

Effort, also present in the SEEV model previously presented, can be divided into three components: “effort competition, effort demands and effort investment” (p. 184). Effort competition refers to the extent to which the same cognitive resources are needed for both tasks. Two tasks that both require a high level of cognitive resources are predicted to result in a detriment in performance. Effort demands bear on the difficulty of the particular tasks for a particular individual. Effort demands, therefore, take into account the cognitive ability and skills of an individual, such that a task of increased difficulty can be offset by an individual of higher
skill or cognitive capacity, or an individual that has extensive practice in a particular task such that aspects of the task have been automated. The last component, effort investment, involves the degree to which a particular individual performs a task to the best of her ability (i.e., percentage of cognitive resources invested). The component of effort, therefore, reflects both the motivation of the individual to perform the task, and the difficulty of the tasks for the individual, a concept initially proposed by Kahneman (1973). In second language acquisition, the importance of these aspects of effort have been explored in studies on motivation (e.g., Ellis, 2008; Gardner & MacIntyre, 1991), and the task-induced involvement load (Hulstijn & Laufer, 2001; Keating, 2008; Laufer & Hulstijn, 2001).

In SLA, motivation is seen as both the desire to do well and the effort undertaken by the language learner. Various types of motivation are identified in second language acquisition, including integrative motivation, linguistic self-confidence as motivation, attributions, intrinsic motivation, and instrumental motivation (Ellis, 2008). One study on instrumental motivation (Gardner & MacIntyre, 1991) in which a financial incentive was given to participants for performing well on a task, is most relevant to the current discussion of the “effort invested” (p. 183) component of effort, given that a financial incentive would likely increase the amount of effort invested. In this study, ninety-two college students that studied French in high school, but were not currently studying French, participated. The participants were randomly assigned to one of two experimental conditions, only one of which included a 10-dollar incentive to perform well on a French vocabulary-learning task. Gardner and MacIntyre (1991) found that the students provided with this incentive performed significantly better than those who did not have this incentive, showing that the incentive motivated students to increase their effort investment. Following Wickens’ (2002) Multiple Resource model, the higher effort investment motivated by
the incentive would likely also make the task more likely to be successfully performed under multitasking conditions. Laufer and Hulstijn’s (2001) Task-Induced Involvement Load construct is relevant to the remaining aspect of effort in Wickens’ (2002) model, namely, effort demands.

The Task-Induced Involvement Load construct indexes the processing load of vocabulary learning tasks. According to the construct (described in more detail under Empirical Evidence & Depth of Processing in the L2), three dimensions are important in determining a particular vocabulary learning tasks’ involvement load: need, search, and evaluation. Need refers to the motivation of a learner to discover the word’s meaning, search refers to the investigation into the word’s meaning, and evaluation refers to the relevant analysis of the word and its definition, synonyms, context, and/or translation. Those tasks that are greatest in the dimensions of need, search, and evaluation have the highest involvement load, similar to effort demands in Wickens’ Multiple Resource Model (2002) model. Following Wicken’s model (2002), tasks with higher involvement load or effort demands would be predicted to be more difficult to perform simultaneously with another task, especially if the additional task also requires substantial effort.

Qualitative resource similarity, the second component in determining the impact on performance of dual-tasking, refers to the degree to which two tasks involve the use of the same types of resources. The more similar the resources needed for the two tasks, the higher the amount of interference, and, in turn, the greater the negative impact on performance. Wickens (2007) identifies three dimensions in which tasks may be similar or different: (1) Processing modality; (2) Processing codes; and (3) Processing stages. Processing modality refers to whether a task is visual or auditory. According to Wickens (2007), it is less likely that there will be a detriment in performance if the tasks are of distinct modalities. In second language acquisition, the visual modality is expressed as reading and writing, while the auditory modality is expressed
in listening and speaking. Wickens (2007) points out evidence of the role of modality in a
summary of studies on attention to subtitles (d’Ydewalle, 2002) and a study on dual-tasking
involving shadowing (Rollins & Hendriks, 1980).

D’Ydewalle (2002) cites studies in which participants were able to simultaneously
process the soundtrack and written subtitles while watching movies (d’Ydewalle & Pavakanun,
1995, 1997; d’Ydewalle, Praet, Verfaillie, & Van Rensbergen, 1991). D’Ydewalle and
Pavakanun (1995) demonstrated that comprehension is not negatively affected by subtitles. In a
later study (1997), they found that participants were able to learn from listening to the soundtrack
and reading the subtitles, when one of the two was in a foreign language, indicating that they
were able to pay attention to both the soundtrack and the subtitles. In addition, in two similar
experiments, d’Ydewalle et al. (1991) specifically compared the use of same-language subtitles
among native speakers with and without a soundtrack in order to determine if there was a trade-
off between paying attention to subtitles and paying attention to the soundtrack. With regards to
time spent looking at the subtitles, they only found a marginally significant difference between
the condition with the soundtrack and the condition without the soundtrack, indicating that any
tradeoff between listening to the soundtrack and reading the subtitles is likely minimal. The
ability of participants to pay attention to the two modalities with relative ease provides evidence
in favor of Wickens’ (2007) postulation that there is less interference between two tasks of
different modalities. While it might be pointed out that d’Ydewalle (2002, p. 60) also cites a
study that found that participants were also able “to switch effortlessly between the visual image
and the subtitle” (d’Ydewalle, Van Rensbergen, & Pollet, 1987), the visual image was not
language-related, and, therefore, of a different processing code, which Wickens (2007) also
postulates to decrease interference.
Rollins and Hendriks (1980) explored the role of modality in three separate word processing experiments. In the first two of the three experiments, participants were presented with target words under auditory and visual modalities, and categories and rhymes visually. They were then asked to shadow a series of digits or numbers until they heard or saw the target. Of the two modalities, Rollins and Hendriks (1980) found that participants were most successful when the word targets were presented visually. Given that participants shadowed the digits or words under the auditory modality, the higher performance under the visual method of presentation of the target words can be interpreted as a reflection of the greater ease of processing for two tasks of different modalities, as predicted by Wickens (2007). In the third experiment, which required participants to name the categories of the words presented to them, fewer errors were again found for the visual condition, providing further evidence in favor of Wickens’ (2007) Multiple Resource Model.

Processing codes, the second component of qualitative resource similarity, touches on whether the task involves verbal/linguistic material or spatial/analog non-verbal material (p. 186). According to Wickens (2007), it is easier to perform two tasks simultaneously if they involve two different codes, meaning that one is verbal and the other is not verbal. Horrey and Wickens (2004), for example, found that drivers were able to simultaneously drive, which can be characterized as a spatial code task, and perform an in-video technology device number activity, a verbal/linguistic material task. As one of the processing codes is “verbal/linguistic material,” dual language tasks would always involve two tasks of the same processing code, making this particular category of qualitative resource similarity less relevant to second language acquisition.

The third component determining qualitative resource similarity, processing stages, distinguishes between tasks involving perception and working memory capacity, and those
involving execution and selection of tasks. With regards to second language acquisition, Wickens (2007) relates the working memory capacity and perception tasks to comprehension, and execution and selection to production. Such a distinction is most clearly made in Gass and Selinker’s (2008) model of the second language learning process, in which comprehension and output are on opposite ends of the learning spectrum.

The third factor of the Multiple Resource Model, allocation, refers to the relative amount of cognitive resources devoted to each of the two tasks being performed (Wickens, 2007). Allocation determines which of the two tasks are negatively affected by the dual-tasking in the event that insufficient cognitive resources are available to undertake both tasks at 100%.

Wickens (2007) argues that the SEEV model can be used to predict the allocation of resources, with tasks higher in saliency, expectancy, and value, and lower in effort more likely to be allocated the necessary resources.

In the context of the current dissertation, which seeks to explore the ability of participants to process for both form and meaning while reading a text, the prediction from Wicken’s (2007) models, as well as all but one (Treisman, 1964) of the attentional models explored here (Broadbent, 1958; Carbonell, 1966; Kahneman, 1973; Senders, 1964), is that performance will necessarily diminish as attentional resources are exhausted. The effort factor suggests that that point will not be the same for all individuals, as some may have superior cognitive abilities. This possibility is the motivation for the inclusion of working memory capacity as an individual difference variable in the current thesis. The effort factor also suggests that the difficulty of each of the two tasks, in this case processing for form and processing for meaning, will also play a role. Hence, the current experiment explores two levels of difficulty, in the form of processing levels, to examine this possibility as well. A minimum performance for processing for form is
also required, given the need to allocate resources to this activity, such that the deficit would be seen in terms of reading comprehension. While cognitive models of attention provide support and motivation for the current thesis, a more direct connection between the current thesis and attention can be derived from models of attention in second language acquisition research.

Models of Attention in SLA

In order to account for the limited-capacity of attention proposed by Broadbent (1958), researchers postulated two modes of information processing: controlled processing and automatic processing (e.g., McLaughlin, 1987; Shiffrin & Schneider, 1977). While the specific accounts of this hypothesis vary, most agree that controlled processing is slow, capacity-limited and conscious. Automatic processing, on the contrary, is viewed as fast, and as having very little or no demand on attentional resources (DeKeyser, 2001; Schneider & Shiffrin, 1977). Specifically, Schneider and Shiffrin (1977), who consider memory a “large and permanent collection of nodes that become complexly and increasing inter-associated and interrelated through learning” (p. 2), see controlled processing as the temporary activation of a sequence of these nodes. Through practice and repetition, they argue, the sequence of nodes becomes automatic and no longer requires processing, freeing resources for other controlled processes. McLaughlin (1987), drawing heavily on Schneider and Shiffrin (1977), argues that SLA can be viewed under this information-processing framework.

Second Language as a Cognitive Skill

McLaughlin (1987) proposes that learning a second language is a cognitive skill that involves both controlled and automatic information processing, and argues that language can be divided into a series of complex tasks, and that these can subsequently be divided into sub-tasks.
Citing Levelt (1978), McLaughlin gives speaking as an example of a complex task that can be divided into a series of basic subtasks such as “retriev[ing] the appropriate lexicon, utiliz[ing] the correct syntactic rules, and meet[ing] pragmatic conventions” (p. 135). Each of these subtasks, McLaughlin argues, is initially undergone via controlled processes, but, with practice, can be automatized. Automatization of these basic subtasks, in turn, frees up attentional resources to focus on higher order goals such as “decid[ing] on a topic and express[ing] a particular intention” (p. 135). As learners gradually automatize subtasks, according to McLaughlin (1987), they also begin to restructure and organize the knowledge they have attained. As pointed out by McLaughlin, Karmiloff-Smith (1986) offers a potential explanation for how this restructuring occurs. According to Karmiloff-Smith (1986), during phase one, learners use external stimuli to produce the desired output. In phase two, they evaluate and organize their internalized representations without processing additional external stimuli. In phase three, learners again use the external stimuli, this time to restructure their internal representations based on the additional data. An interesting aspect of McLaughlin’s (1987) theory is that while Schneider and Shiffrin (1977), and the standard view of this account of information processing, argue that controlled processing is “conscious” (in the sense of awareness), McLaughlin (1987) argues that both controlled and automated processing can be conscious or not. This implies that language learning can occur without awareness, a contention of great debate in second language acquisition, and one that is addressed by Tomlin and Villa’s (1994) model of input processing in SLA, as well as by Robinson (1995) and Schmidt (2001).
Attention as Alertness, Orientation, and Detection

Tomlin and Villa (1994) argue that attention in second language acquisition can be divided into three networks: alertness, orientation, and detection. According to Tomlin and Villa (1994), alertness implies being ready for the stimuli. For instance, students are alert when they are ready and anxious to learn, as opposed to bored and uninterested. Orientation is the direction in which attentional resources are focused. For example, in second language acquisition, orientation could be focused on the meaning of an utterance or on its grammatical forms. Detection takes place when learners cognitively register stimuli, a required condition for processing to occur. One example of detection would be if a learner, either with or without awareness, were to detect that Spanish speakers often omit the grammatical subject when speaking. Citing Posner’s empirical investigations that the three networks each activate different components of the brain (Posner & Petersen, 1990; Posner & Rothbart, 1992), Tomlin and Villa (1994) argue that alertness, orientation, and detection can operate independently of one another. Given this evidence, they also argue that while alertness and orientation increase the chances of detection, and, therefore, further processing, they are not necessary for detection to occur. In addition, they make the argument that detection without awareness can lead to further processing of the input, citing studies that have found detection to occur independently of awareness (e.g., Marcel, 1983, third experiment). Many theorists have cast doubt on Tomlin and Villa’s (1994) claims that input can be processed further without awareness and that detection can occur without orientation or alertness (Simard & Wong, 2001; Robinson, 1995; Schmidt, 2001). Specifically, they argue that Tomlin and Villa (1994) rely on evidence (e.g., Marcel, 1983) that is not specific to second language learning (Simard & Wong, 2001), and that the empirical evidence that they cite, as well as other available empirical evidence, only suggests that
perception without awareness is possible, but not learning without awareness (Simard & Wong, 2001; Robinson, 1995; Schmidt, 2001). At least one study in second language acquisition has specifically addressed and found support for Tomlin and Villa’s (1994) model (Leow, 1998) in relation to his three attentional functions.

**Empirical Evidence**

Leow (1998) employed crossword puzzles in his experimental design to isolate and test the three attentional networks that Tomlin & Villa (1994) postulated. In his study, 83 final participants completed a multiple-choice and a written production pre-test on the target form, one of four Spanish crossword puzzles, and then one post-test and two delayed post-tests on the same assessment tasks on the same target form. Each of the crossword puzzles contained instances of the target form, the Spanish irregular stem-changing verbs in the preterit tense. The attentional network controlling the mere readiness to receive the material, alertness, was operationalized in the first crossword by simply allowing participants to successfully complete the puzzle regardless of whether or not they conjugated the irregular verbs correctly. Orientation was operationalized in the second and third crossword puzzles with boldfaced instructions pointing out to participants that “some of the forms of the verbs are IRREGULAR” (p. 138). Detection, the cognitive registration of stimuli, was operationalized in the third and fourth crossword puzzles by providing clues that provided the vowel that needed to be changed in the stem changing verbs. In this way, participants assigned to the fourth crossword puzzle were given the vowel clues, but not the boldfaced instructions, meaning that their crossword puzzle was operationalized for alertness and detection, but not orientation. Leow (1997) found that both the group that completed the crossword puzzle operationalized for detection, orientation and
alertness and the group that completed the crossword puzzle operationalized for detection and alertness, performed significantly better on the post-tests than the alertness group and the alertness and orientation group. Leow (1998) contended that these results support Tomlin and Villa’s (1994) “fine-grained model of attention” (pp. 147-148) arguing that they suggest that detection plays a “crucial” role in acquisition, and that alertness and orientation are not necessary for detection to be effective. At the same time, Leow also pointed out in his note 4 that “the question of whether awareness is crucial for intake at the level of detection was not an issue” (p. 155).

Simard and Wong (2001), however, counter that methodological concerns regarding the operationalization and isolation of alertness, orientation and detection should be considered in interpreting the results. They note that the absence of the clue that some verbs are irregular does not necessarily imply that participants did not in fact orient themselves to these verbs during the completion of the crossword puzzle, as well as that the clue itself did not necessarily guarantee orientation. They also point out that alertness is present in all four crossword puzzles, as identified in Leow’s (1998) Table 1 (see page 138). This table indicated whether alertness, orientation, and/or detection was involved in each crossword.

Therefore, the isolation of the three variables may have been compromised, which arguably lessens the evidence in support of Tomlin and Villa’s (1994) model that these three networks do indeed operate independently (cf. Leow, 2002 for a rebuttal). While Schmidt (2001) does not separate detection from other aspects of attention, he does attribute great importance to both attention and awareness in the process of intake.
The Noticing Hypothesis

According to Schmidt’s noticing hypothesis (1990), L2 learners must minimally perceive and be aware of an element of the target language in order to potentially process it for intake. Schmidt differentiates between three degrees of awareness: perception, noticing, and understanding. He explains that perception, as it is commonly understood, “implies mental organization” and “creat[ing] internal representations of external events” (p. 132). Perception without awareness has been found to occur in the form of preferring words graphemically semantically related to words processed without awareness (Marcel, 1983), and preferring the spelling of homophones that the context of the phrases processed without awareness suggested (Eich, 1984). Schmidt (1990, 2001) differs from Tomlin and Villa (1994) in that he claims that instances of subliminal perception, such as those demonstrated in these studies (Eich, 1984; Marcel, 1983), will not lead to further processing. For further processing to occur, Schmidt (1990) argues that, minimally, a learner must notice the information. Noticing requires both attention and a low level of awareness. According to this hypothesis, therefore, attention and awareness are intricately connected in their relationship to learning. In support of his hypothesis that noticing is required for intake to occur, Schmidt (1990) cites a journal study in which no evidence of unaware learning was found (Schmidt & Frota, 1986).

Empirical Evidence

Schmidt and Frota (1986) detail the linguistic experiences of the former researcher, referred to as R in the study, while he was learning Portuguese during his 5-month stay in Rio de Janeiro, Brazil. Prior to traveling to Rio de Janeiro, R had only minimal knowledge of Portuguese, gained by listening to a Portuguese informant giving grammatical judgments. The researchers divide R’s learning experience in Rio de Janeiro into three stages during which the
researcher wrote extensively, although not systematically, about his linguistic experience in the form of a diary. R interacted with native speakers primarily in stages 2 (5 weeks) and 3 (14 weeks), and received instruction only in stage 2, describing his experience in stage 1 as only involving very minimal interaction in Portuguese. The researchers recorded R’s conversations with a native speaker of Portuguese, S, to measure R’s output and S’s input. In the analysis of R’s notes, they found that R did not mention noticing the compound perfect tense or the compound past perfect, despite both being present in the input, and only mentioned noticing the inflected future towards the end of stage 3. This absence of noticing coincided with R’s output of the three forms, as, of the three forms, R only produced the inflected future, and produced it precisely towards the end of stage three. In addition, many other forms that were noticed were extensively used in R’s output. Schmidt and Frota (1986) suggest, from this data, that noticing appeared to have been necessary for R to learn and produce Portuguese verb forms, thus providing evidence in favor of Schmidt’s (1990, 2001) noticing hypothesis. In addition, Tomlin and Villa (1994) assert that the diary study (Schmidt & Frota, 1986) may have been too coarse grained in that the act of noticing and writing about it in the diary were not concurrent. In a response to this criticism, Schmidt (2001:19) cites a more recent study that measured awareness using an online method of data collection (Leow, 1997).

Leow (1997) examined the impact of awareness on recognition and production of the target form. In the experiment, 28 participants completed a crossword puzzle with instances of irregular stem-changing Spanish verbs, the target form of the study, as well as clues which intersected with the target form and, in so doing, revealed the irregular nature of the target form (as in Leow’s 1998 detection group). Participants also thought aloud while completing the crossword puzzles. In the analysis of the think-alouds, Leow (1997) identified three levels of
awareness among participants: (1) A lower level of awareness only showing evidence of a cognitive change; (2) A middle level of awareness exhibiting a cognitive change and meta-awareness but no verbalization of the underlying rule; and (3) A high level of awareness exhibiting a cognitive change, meta-awareness together with verbalization of the underlying rule. Leow (1997) found that as level of awareness rose, so too did the performances on the two assessment tasks.

Given that participants that verbalized the morphological rule were coded as having a high level of awareness in the study, with those that merely noticed the difference in forms being labeled as having low levels of awareness, depth of processing and degree of awareness, as operationalized in this study, can be understood as related concepts. Therefore, Leow’s (1997) findings suggest that the quality of noticing, or level of awareness, plays a role in processing, providing indirect support to the noticing hypothesis. Later studies have provided support for the researcher’s finding that higher levels of awareness lead to better performance in production and recognition tasks on the target form (e.g., Rosa & Leow, 2004; Rosa & O’Neill, 1999).

Nonetheless, given that Leow (1997) did not make a comparison with learners that did not notice the items to any degree, or, in other words, were completely unaware, the study falls short of demonstrating the central premise of the noticing hypothesis, namely, that intake without awareness cannot occur. Recent studies examining whether or not awareness (comparable to noticing given that noticing implies a low level of awareness), in addition to attention, is necessary for intake to occur have provided indirect evidence both in favor (Faretta-Stutenberg & Morgan-Short, 2011; Hama & Leow, 2010; Leow, 2000; Rosa & Leow, 2004), and against (Chan & Leung, 2013; Chen, Guo, Tang, Zhu, Yang, & Dienes, 2011; Leung & Williams, 2011, 2012, 2014; Williams, 2004, 2005) the noticing hypothesis.
Rosa and O’Neill (1999) examined the impact of awareness on the intake of conditional sentences in Spanish. In the study, 67 native speakers of English in fourth semester Spanish classes thought aloud while they put together 10 jigsaw puzzles, with the pieces that fit together in each puzzle being the two clauses of a conditional sentence. Rosa and O’Neill (1999) depended on the concurrent verbal reports to code for awareness, which they divided into three levels: understanding; noticing; and no verbal report. While Rosa and O’Neill (1999) found those participants that were coded for awareness at the level of understanding to perform significantly better than the two remaining groups, no difference was found between the noticing group and the no verbal report group. In addition, participants at each of the three levels of awareness improved from the pre-test to the post-test indicating that awareness may not be necessary for intake. Notably, however, the improvement for the no verbal report group was highly variable across participants.

Rosa and Leow (2004) expanded on Rosa and O’Neill’s (1999) jigsaw puzzle on awareness. In their study, however, the jigsaw puzzle was computerized, and, instead of the condition requiring participants to search for a rule in Rosa and O’Neill (1999), participants in Rosa and Leow (2004) were assigned to two different types of feedback that varied by degree of explicitness. In addition, participants in Rosa and Leow (2004) also completed a questionnaire that was used, in conjunction with the Think-Aloud protocols, to code for awareness. Unlike in Rosa and O’Neill (1999), Rosa and Leow (2004) did find a significant difference between the noticing awareness group and the no report awareness group, and found that the all report groups improved from the pre-test to the post-test. Given that the statistical analyses employed in the two studies (Rosa & Leow, 2004; Rosa & O’Neill, 1999) factored out the influence of the remaining conditions, which differed in the two studies, it is unlikely that the difference in
findings can be attributed to them. A more plausible explanation for the difference in the two studies, particularly given the variability in improvement for the no verbal report group in Rosa and O’Neill, is that Rosa and Leow’s (2004) more exhaustive analysis of awareness made it much less likely that participants would be miscoded. This may also explicate the results obtained in other studies with varying results (e.g., Chen et al., 2011; Faretta-Sutenberg & Morgan-Short, 2011; Hama & Leow, 2010; Leow, 2000; Leung & Williams, 2011, 2012, 2014; Williams, 2005).

Leow (2000) extends Leow (1997) by coding for an unaware group and employing an additional offline measure of awareness. The target structure was, following Leow (1997), Spanish irregular stem-changing verbs in the preterit. In the experiment, 32 final participants were instructed to complete a pre-test on the target forms, to think aloud while they completed a crossword puzzle with instances of the target form, and, lastly, to complete a post-test on the target forms. The vowels crucial to the irregularity of the target forms were already completed in the crossword puzzle, as opposed to being filled in by the participants with the answers to other clues as in the original study, in order to avoid promoting noticing and maintain a maximum number of unaware participants. Leow (2000) found that awareness, as determined by probe questions and the concurrent verbal protocols, was crucial to performance, as aware participants improved from the pre-test to the post-test, while unaware participants did not. Williams (2005) notes, however, that a key limitation in Leow (2000) is that it concerns discrete item learning, and, consequently, does not address the potential of unaware participants learning to generalize based on a pattern to which they are exposed.

Williams (2005), a methodological improvement of Williams (2004), empirically examined the ability of unaware participants to learn a linguistic generalization. The target form
of the experiment consisted of four determiners that indicated both the distance and the animacy of the modified noun. Forty-one participants first completed a vocabulary pretraining exercise in which they learned to distinguish the determiners in terms of their distance meaning. Subsequently, in the training task, participants listened to and repeated a series of sentences with these determiners from which the participants could theoretically generalize their animacy values. Participants then completed a test phase in which they had to correctly select the endings of a series of sentences with two options, one with an animate determiner and one with an inanimate determiner. At the completion of the test phase, participants were interviewed to determine whether or not they were aware of the animacy distinction. As expected, aware participants performed very well in the test, averaging 91.4% or above chance with an alpha value of .001. Contrary to Leow (2000), Williams (2005) also found that unaware participants performed at a level significantly above chance. Williams (2005) argues that this last result provides evidence that form-meaning connections can be made without awareness. Hama and Leow (2010), nonetheless, identify, among other issues, a key methodological concern with their offline operationalization of unawareness, indicating that the evidence found in Williams (2005) with regards to learning without awareness should be interpreted with caution.

With regards to the methodology of the experiment, Williams (2005) relied on a post-test interview to determine awareness. Specifically, he asked participants “what criteria they had used to make their choices” as to which determiner article to choose to complete the test phase sentences (p. 283). As Hama and Leow (2010) explain, however, this type of questioning may be insufficient to detect low levels of awareness. Additionally, potentially due to memory loss, some participants may have been aware during the experiment but simply did not report being so after the experiment. The potential coding of aware participants as unaware is sufficient to
question Williams’ (2005) results with regards to the group coded as unaware. In order to address this issue, Hama and Leow (2010) employed concurrent verbal protocols in their follow-up to Williams (2005).

Hama and Leow (2010) extended Williams (2005) by employing a triangulation of methods to code for awareness. Similar to Williams (2005), in Hama and Leow (2010) participants, 34 in total, completed a vocabulary pretraining task, a training task on the four determiners, a multiple-choice assessment task on the target form, and subsequently answered questions designed to probe their awareness. Unlike Williams (2005), however, participants in Hama and Leow (2010) completed a production assessment task in addition to the multiple-choice assessment task, and Hama and Leow’s (2010) multiple choice assessment task included four options, as opposed to Williams’ (2005) two. Arguably the most noteworthy distinctions between the two studies, however, is that participants in Hama and Leow (2010) thought aloud while completing the tasks, which provided the research with an additional tool to code for awareness. Indeed, while evidence of awareness from the offline and online measurements for the most part coincide with each other, with eight participants being determined as aware by both, and only three participants showing evidence in just one of the two measurements (two only in the online measurement, and one only in the offline measurement). This potentially may explain why Hama and Leow (2010), in contrast to Williams (2005), found no evidence of learning among the unaware participants. Another potential difference that may have had an influence on the difference in results, as Faretta-Sutenberg and Morgan-Short (2011) note, is the varied language backgrounds of the participants in Williams (2005).

Faretta-Sutenberg and Morgan-Short (2011) replicated Williams (2005) with a different participant base in order to explore the role of participant language backgrounds on implicit
learning. Thirty undergraduate native speakers of English took part in the study. Of these participants, 14 participants spoke a gendered second language at least at an intermediate level. The methodology closely followed Williams (2005) in order to allow a fair comparison of the results. Faretta-Sutenberg and Morgan-Short (2011) coded for awareness both as in Williams (2005), where participants who mentioned animacy were coded as aware and those who did not were coded as unaware, and as in Hama and Leow (2010, detailed above), in which participants were coded as No Report, Aware at the Level of Noticing, and Aware at the Level of Understanding. Under Williams’ (2005) coding scheme, neither the aware nor unaware participants performed above chance for the generalization items. Under Hama and Leow’s (2010) coding scheme, however, participants who were coded as aware at the level of understanding did perform above chance for the generalization items. Faretta-Sutenberg and Morgan-Short (2011) also looked at the impact of language backgrounds. They found no significant difference for generalization items between the performance of unaware participants that spoke an L2 and those that did not. Years of education was found to have a significant relationship with performance on generalization items for aware, but not unaware participants. Faretta-Sutenberg and Morgan-Short (2011) suggest that, despite the lack of any significant difference for any of the individual variables measured in this study in terms of the performance of unaware learners, the difference in findings with regards to unaware learning between the two studies may still be attributable to individual differences, perhaps those not explored in Faretta-Sutenberg and Morgan-Short (2011). While this may indeed be the case, an alternative explanation may be that the extent to which the offline measurement used in the two studies was successful may have varied between the two studies in terms of providing evidence of participant
awareness (i.e., Faretta-Sutenberg & Morgan-Short may have had less instances of aware participants not mentioning their awareness in the survey).

Chen et al. (2011) extended Williams (2005) by examining awareness on a trial-by-trial basis in three separate experiments. In the first experiment, 40 undergraduates and graduates whose native language was Chinese and spoke English as a second language participated in the study. Like Williams (2005), except in Chinese characters, four determiners, which determined the animacy and distance of the noun they modified, were used. Participants were told about the distance, but not the animacy distinction. In the training phase, participants read 48 sentences embedded with the determiners aloud, and, after each sentence, indicated whether the noun modified by the determiner was near or far. The testing phase consisted of a sentence completion task in which participants chose one of two options. After each choice, the participants were given twenty seconds to give one out of four reasons for their choice: (1) guess; (2) intuition; (3) memory; or (4) rule. Those responses that corresponded to reasons (1) and (2) were combined as representing “unconscious structural knowledge” (i.e., unawareness, p. 1754). Chen et al. (2011) found that these responses were correct at a rate significantly higher than chance both for trained and for generalization items. Chen et al. (2011) also combined responses corresponding to reasons (3) and (4). These responses were also correct at a rate significantly higher than chance for both types of items. Interestingly, no participant mentioned any rule in the post-task debriefing. Concerned that participants may have been relying on pairing patterns, as opposed to the animacy distinction, Chen et al. (2011) conducted a second experiment with novel nouns in the testing phase and found the same results with regards to unconscious structural knowledge, but found that those responses corresponding to memory or rule reasons were correct at a rate significantly above chance for trained but not generalization items. In the third experiment, Chen
et al. (2011) substituted the linguistically relevant animacy distinction for an arbitrary distinction, namely animal size (large vs. small). In this experiment, accuracy was not found to be above chance for either trained or generalization items. Based on these three experiments, Chen et al. (2011) concluded that implicit learning connections between form and meaning are possible when these connections are linguistically relevant.

Chen et al.’s (2011) measurement of awareness, however, is a potential methodological concern for the study. According to the sensitivity criterion, “a measure of awareness should be sensitive to all of the conscious knowledge that could have guided learning…” (Robinson, Mackey, Gass, Smith, 2012, p. 19). Arguably, the subjective measurement used in Chen et al. (2011) to determine awareness, which required participants to select from one of four options, is not sufficiently fine-grained to be sensitive to “all of the conscious knowledge that could have guided” the participant in the sentence completion task (p. 19). One might speculate, for example, that a participant that feels that she only has limited grasp of a rule might select “guessing” or “rule” as a reason for her response. The exclusive use of this measurement of awareness is particularly concerning in light of the fact that it was not corroborated by the post-test debriefings, where not a single participant mentioned the rule, despite all but three participants having cited “rule” or “memory” as a reason for their response. In addition, one key question is left unanswered in Chen et al.’s (2011) reporting of the data: Chen et al. (2011) does not mention whether or not there were instances in which participants selected intuition or guess after they had already selected rule, an occurrence which would clearly question the implicit nature of their learning in the former cases.

In an innovative approach to the issue of unaware learning, Leung and Williams (2011) employed reaction times in an attempt to measure implicit learning. The target forms of the
current experiment were four articles that indicate whether the noun they modify is an adult or a child, as well as whether the noun is the agent or the patient of the action. In the experiment, 25 participants, as in previous awareness studies (Hama & Leow, 2010; Williams, 2005) first completed a vocabulary pretraining exercise designed to introduce them to one aspect of the determiners meaning, which, in the case of the current experiment, was the adult versus child distinction. Following this exercise, participants were shown 120 pictures with two individuals, and, for each picture, were required to: (1) describe what they saw; (2) listen to an audio description with the target form, but not necessarily in standard English SVO order; (3) identify the side of the picture containing the individual named in the description; and (4) reformulate the sentences into correct English. The last 32 pictures formed the testing portion of the data, with 16 pictures following the same pattern serving as the control block, and the remaining 16 reversed the patient/agent mapping of the determiners. At the end of the experiment, participants were asked if they “had any feelings about when gi versus ro and ul versus ne were used” (p. 45). Answers to this last question were used to determine whether each participant was aware or unaware. In the qualitative analysis, Leung and Williams (2011) determined that five of the 25 participants were aware by the end of the experiment. Among the 20 participants coded as unaware, the reaction times during the violation block were significantly different from those of the control block. The authors suggested that this indicates that the participants learned the form-meaning connections for the determiners, and, consequently, that learning can occur without awareness.

In two experiments similar to Leung and Williams (2011), Leung and Williams (2012) looked at both reaction times and error rates to explore implicit learning. Thirty-three native English speakers, many of who spoke one or multiple foreign languages, participated in the first
experiment. Four determiners, *ge, ul, ro,* and *ne,* which indicated the distance and the animacy of
the noun they modified, were used. Participants, however, were only informed of the distance
distinction. Participants were exposed to 344 slides which were divided into two frames with one
object in each frame that was either far or near, and either animate or inanimate. After each slide,
participants heard a noun phrase, which corresponded to one of the two objects, and had to
determine if the phrase referred to a living thing or a non-living thing. The training phase
consisted of 136 slides for which participants received feedback indicating whether or not they
chose correctly, and 136 slides for which they did not receive feedback. The training phase was
followed by 8 transition slides, 32 control slides, and 32 violation slides, the last of which
reversed the articles in the audio. After the slide presentation, participants took part in a post-
experiment interview in which they were asked what they thought the experiment was about, if
they noticed anything odd towards the end of the experiment, and “if they had any feelings about
the different conditions in which *gi* versus *ul* and *ro* versus *ne*” were used (p. 13). From the
analysis of the postexperiment interview, Leung and Williams (2012) coded 20 of the 33
participants as unaware because they “claimed not to have been aware of the relationship
between articles and animacy during the experience or to have noticed anything unusual towards
the end” (p. 14). Leung and Williams (2012), note, however, that 7 of these participants did
guess animacy, but were not able to state which articles had animate and inanimate values. For
the unaware participants, Leung and Williams (2012) found a significant increase in reaction
times and error rate from the control to the violation slides. Leung and Williams (2012) also
found a significant increase in reaction time and error rate for 12 of the 13 in the unaware group
that did not guess animacy. For participants coded as aware, while a significant increase was
found for reaction time between the control and the violation phase, no significant difference in error rate was found between the two phases.

Different results were found in the second experiment, which varied from the first only in that size was used instead of distance, and that participants were informed about the animacy distinction, not the size distinction. In the experiment, 20 participants were coded as unaware, 9 of which guessed size as the differentiating value. For the unaware participants, no significant difference was found between the control and the violation blocks for error type or reaction time. In addition, no significant difference for reaction time was found for the 11 participants in the unaware group that were not able to guess correctly. A significant difference was found, however, for the 9 participants that did guess size as being the remaining distinction between the four determiners. With regard to the aware group, no significant difference was found for error rate or reaction times.

Leung and Williams (2012) argue that the results of the two experiments taken together provide evidence in favor of implicit learning of meaning-form connections. They also contend that the results show that the nature of the unknown values, animacy in experiment 1 and size in experiment 2, may differentially constrain or facilitate learning. One potential reason they provide for why animacy may facilitate implicit learning is that animacy impacts the grammatical structures that are used in the participants’ native language, English.

In their most recent study, Leung and Williams (2014) explored the importance of L1 linguistic knowledge in implicit learning in three separate experiments. Four determiners were used in each experiment, but participants were only told that they could be distinguished in terms of distance (near/far). In the first experiment, the determiners could be categorically divided by both distance and animacy. In this experiment, twenty-seven native speakers of Cantonese
Chinese and 30 native speakers of English were exposed to a series of trials of noun phrases, which included the four determiners. The English participants were presented with a series of noun phrases, which could be divided into four blocks. In the first three blocks, participants were shown numerous correct determiner-noun combinations and were asked to first indicate the animacy meaning, and to second indicate the distance meaning of the combinations. When they clicked on the key corresponding to the incorrect answer, feedback was provided in the form of an unchanged display. The first half of the fourth block were the control trials, and the second half were the violation trials, in which the opposite animacy and distance articles were used. The Chinese version was similar, except that more grammatical trials were presented in the training phase, and, in the testing block, sixty grammatical fillers were used. Following the presentation of determiner-noun pairs, participants filled out an awareness questionnaire. The questionnaire first probed their knowledge of the animacy rule, and then allowed them to guess. This led to two levels of awareness, participants that were not deemed aware based on the former (unaware participants), and those that did not even guess the animacy rule (henceforth, non-guessers). For the English participants, the authors found that a significant difference between the control and violation trials in terms of response time for the unaware participants, but not for the non-guessers, while both were found to be significant for the Chinese participants. The second experiment was similar to the first, except that the hidden meaning was changed to “number of capital letters” for the English participants, and “number of strokes” for the Chinese participants and fewer blocks of trials were used (p. 14). No significant differences were found between the control and violation blocks for this experiment. Finally, in the third experiment, also similar to the first except that the unknown regularity was a long/flat distinction, which is present in Chinese but not in English, differences between the control and the violation trials in terms of
reaction time were found for the Chinese group but not for the English group. Leung and Williams (2014) suggest that the results of this experiment both suggest that unaware learning is possible in second language acquisition, and that it may depend on aspects of one’s native language.

Three key methodological concerns, however, must be taken into consideration when interpreting Leung and Williams (2011, 2012, 2014). First, in each of the three studies, as in Williams (2005), Leung and Williams inquired on their participants’ awareness after they completed the experiment. As Leow and Hama (2013) point out, however, probing for awareness at this stage may be insufficient to ensure proper coding of participants as aware and unaware, due to the possibility of veridicality. The potential coding of aware participants as unaware could have skewed Leung and Williams’ (2011, 2012, 2014) results with regards to the group coded as unaware. Second, in Leung and Williams (2011, 2012), the inquiry involved asking participants “if they had any feelings” as to the usage of the different determiners (2011, p. 45). This particular question is only likely to solicit awareness at the level of understanding (Leow and Hama, 2013), however, and, as Leow (1997) demonstrates, awareness can also exist at the level of noticing and reporting. Therefore, Leung and Williams (2011, 2012) should qualify their claims by noting that they are referring to participants without awareness at the level of understanding. Leung and Williams (2014), unfortunately, do not offer sufficient detail with regard to the questionnaire that they used to probe for awareness to be critiqued. A final concern with Leung and Williams (2011, 2012, 2014) is that the aware group is not statistically compared to the unaware group, and, therefore, Leung and Williams (2011, 2012, 2014) cannot demonstrate that a different type of learning, one that is implicit and unaware, has taken place.
In summary, among those studies exploring the ability of language learners to learn without awareness, five claim to have found evidence of learning without awareness (Chen et al., 2011; Leung & Williams, 2011, 2012, 2014; Williams, 2005). The first, Williams (2005), found that learners coded as unaware were able to identify the individual described by the target form at a statistically higher level than chance. Both a replication of this study with a different participant base (Faretta-Sutenberg & Morgan-Short, 2011) and extended version of this study with an online measurement of awareness to reduce the chances of an aware participant being coded as unaware (Hama & Leow, 2010), however, found contradicting evidence. The second, Leung and Williams (2011), also did not take sufficient measures to ensure that no aware participants were coded as unaware, and, in addition, was unable to make a comparison to the aware participants’ reaction times, and claimed that implicit learning is possible based on a 1/10 of a second reaction time difference that is not corroborated by the accuracy data. The third and forth, Leung and Williams (2014) and Leung and Williams (2012), as in Leung and Williams (2011), relied on an offline method alone to measure awareness. The fifth, Chen et al. (2011) relies on exclusively on a coarse-grained subjective measurement of awareness whose data was not corroborated by the post-task debriefing performed. Two of three studies negating the existence of learning without awareness (Hama & Leow, 2010; Leow, 2000), on the other hand, used online and offline measures to measure awareness (Hama & Leow, 2010; Leow, 2000), and, are, therefore, arguably more reliable (Hama & Leow, 2010). Even Williams (2005), who argues in favor of the possibility of awareness without learning, admitted that Leow (2000) provides convincing evidence that item learning is not possible without awareness. Overall, then, the empirical evidence on awareness supports the contention of Schmidt (2001) that both attention and awareness are necessary for intake to occur.
Attention and Memory

Robinson (1995) combines elements of Tomlin and Villa’s (1994) model of input processing in SLA with elements from Schmidt’s Noticing Hypothesis, and posits roles for short-term and long-term memory. Robinson (1995) views detection (Tomlin & Villa, 1994) and noticing (Schmidt, 1990, 2001) as two related concepts, differing only in that noticing but not detection implies some level of awareness. Robinson (1995) claims that detection can occur without awareness, but that this low “quality of attention” is not likely to have a sufficient impact on memory to be useful for language learning (p. 306). Robinson (1995) asserts that the two theories are “reconcil[able]” if noticing is defined as “detection plus rehearsal in short-term memory” (p. 296). According to Robinson (1995), there is a minimal amount of activation that needs to occur in short-term memory in order for a learner to be considered aware of the information. For this reason, once detected, a learner must then allocate sufficient attentional resources to the information in short-term memory in order to notice it. If one considers that this minimum amount of attention is also needed for awareness, Schmidt’s noticing hypothesis can be seen as consistent with Robinson’s (1995) claim. It is in this connection between these two theories that the theoretical importance of awareness is made most clear. If awareness were not necessary, small amounts of attentional resources could be dispersed across multiple tasks, and cognitive overload would not be expected.

Robinson (1995) differentiates between short-term and long-term memory by noting that short-term memory is the activated portion of long-term memory nodes. Additionally, as Robinson (1995) claims, after information is noticed, further activation and rehearsal in short-term memory is necessary for the information to become part of long-term memory. The steps involved in the processing of information in the L2 have been the subject of a number of models
in second language acquisition (Chaudron, 1985; Gass, 1988; Gass & Selinker, 1994; 2008). The theoretical underpinnings of information processing will be discussed in the next chapter.
CHAPTER 2: INPUT PROCESSING

Introduction

Input processing, or the processing of linguistic data available to the learner, is an important strand of research in second language acquisition as it is has been argued to be the first of a series of processes which leads to the acquisition of language (Gass & Selinker, 1994; VanPatten, 2004). Input processing has been extensively explored on both a theoretical and an empirical basis (e.g., Chaudron, 1985; Gass & Selinker, 2008; Pieneman 2007; VanPatten, 1990, 2004; Wong, 2001).

Theoretically, input processing has been described as a “mediating process” between input and a learner’s internalized grammar (Chaudron, 1985, p. 1). Most researchers indeed agree with this general assertion, although some theorize that additional processes take place between input processing and a learner’s internalized grammar (Gass & Selinker, 1994). In addition, Gass and Selinker (1994) assert that input processing can be divided into two types, which they refer to as apperceived input and comprehended input (pp. 296-297). It has also been proposed that input leads to two different types of intake, namely, communication intake and learning intake (Fæerch & Kasper, 1980). Of particular interest to second language acquisition is the nature of the limits to input processing, given its potential ability to inform the teaching profession on the effectiveness of complex tasks.

The potential limits to input processing have been explored on a theoretical basis by VanPatten (2004) and Pienemann (2007). Pienemann (2007) theorized that input processing is constrained by the knowledge by the learner. VanPatten (2004), on the other hand, argued that input processing is limited by the cognitive resources available to the learner. According to
VanPatten (2004), there is a point at which learners become cognitively overloaded, and, therefore, are unable to fully process the input. This necessarily leads the learner, and more specifically at the early stages of language learning, to process less for either the form or meaning of the input, or potentially both.

The potential impact of limited cognitive resources on input processing has been extensively and empirically explored, with varying results (Hulstijn & Hulstijn, 1984; Hulstijn, 1989; Greenslade et al., 1999; Leow, Hsieh, & Moreno, 2008; Morgan-Short, Heil, Botero-Moriarty, & Ebert, 2012, VanPatten, 1990; Wong, 2001). The variation in results may have been due to methodological differences between the studies, as some did not employ a time-limit (Leow et al., 2008; Morgan-Short et al. 2012; and, for the form and meaning condition, Hulstijn, 1989), while others either did not or inadequately operationalized for either processing for form or meaning (Greenslade et al., 1999; VanPatten, 1990; Wong, 2001, as pointed out in Leow et al., 2008), or both (Hulstijn, 1989, form and meaning condition). In addition, the operationalization of meaning in Leow et al. (2008), which employed Think-Aloud protocols, did not control for reactivity, and Morgan-Short et al. (2012), who controlled for reactivity, controlled for rereading in the Think-Aloud group, but not in the Non-Think-Aloud group. Another potential reason for the variation in results, given the theorized importance of cognitive resources in processing for meaning and form simultaneously (VanPatten, 2004), is that working memory capacity may have played a role.

Models of processing in SLA

According to Chaudron (1985, p. 3), input is the “raw data” or information to which the learner is exposed. When the learner is exposed to input, the learner may or may not perceive the
input. Two aspects play a role in determining whether or not input is perceived by the learner, according to Chaudron (1985, p. 2), who cites Corder (1967): (1) the learner’s “internalized interlanguage”; and (2) the learner’s cognitive abilities. The learner’s cognitive abilities consist of the mental abilities that the learners have to acquire knowledge, while the learner’s internalized interlanguage consists of the grammatical structures that have been acquired by the learner.

Pienemann’s (2007) Processability Theory argues that the learners’ internalized interlanguage has a strong influence on what learners are able to process. Specifically, Pienemann (2007) postulates that learners can only process input when they are developmentally ready to receive it. According to this theory, a series of universal developmental stages for L2 learners’ processors exist, and learners, who move sequentially from one stage to the next, are only able to process input that corresponds to their particular stage. The processing hierarchy, first outlined in Pienemann (1998), includes six stages: “(1) No procedure; (2) Category Procedure; (3) Noun phrase procedure; (4) Verb phrase procedure; (5) Sentence procedure; and (6) Subordinate Clause Procedure” (p. 140). Therefore, in accordance with the Processability Theory, learners at stage (3) will not invert the wh-word in wh-questions, as they have yet to acquire the ability to move grammatical items within the verb phrase, which is not available until stage (4). Likewise, at stage (4), learners will be yet unable to process subject-verb agreement, such as the third-person ‘-s’, as this procedure is acquired at stage (5). At stage (5), in contrast, the learner will be able to process this agreement, potentially taking the information in as intake.

Chaudron (1985, p. 1) postulates the role of intake in information processing as being a “mediating process” between input and the learner’s internalized grammar. According to Chaudron (1985), the distinction between input and intake is that the learner is actively involved
in the case of the latter as he works towards approximating the L2. What is taken in (i.e., becomes intake), depends on both the learner’s internalized grammar and psychological variables, such as the ability to perceive and learn. Chaudron (1983) describes two types of intake, *preliminary* intake, or what the learner has just perceived the input, and *final* intake, or the input that the learner has internalized. Chaudron (1985), who identifies these two types of intake as two ends of the continuum of intake, theorizes three intake processing stages which begin with preliminary intake and end with final intake. The first stage is the initial stage in which input is perceived, while the second stage involves the coding of the meaning of the input into long-term memory. The third stage takes place when the learner uses the linguistic information gleaned from the input to modify his internalized L2 grammar.

Gass (1988) and Gass and Selinker (1994, 2008) provide a full model of the second language learning process, and differentiate their theory from Chaudron, by, among other distinctions, providing a more precise definition of intake and distinguishing between apperceived and comprehended input. Gass and Selinker (2008) postulate two levels of input processing, both of which occur before intake: (1) apperceived input; and (2) comprehended input. Apperception, similar to Schmidt’s (1990) concept of *noticing* is when a learner relates input to prior knowledge. Apperception can occur at a variety of levels, such as at the level of meaning, syntax, or phonology. According to Gass and Selinker (2008), frequency, affective factors, and, as in Chaudron (1985), attention and prior knowledge, are factors that influence whether input is apperceived or not, although these factors are not necessarily independent of one another. The affective variable of motivation, for example, could have an impact on whether or not a particular learner pays attention, or, to use Tomlin and Villa’s (1994) terminology, on whether or not the learner is alert and oriented towards the input. Apperception can be
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

considered a “priming” mechanism for further analysis of the input (Gass & Selinker, 2008, p. 485). The next stage, comprehended input, requires the learner to do the “work to understand” (Gass & Selinker, 2008, p. 484). Comprehension of input is seen as multistaged, in that it can involve comprehension of semantics and syntax, among other elements of the input. The key difference between apperceived and comprehended input is that apperceived input involves the learner noticing that something is occurring in the input, but comprehended input is when the learner discovers what exactly this is. As in the case of apperceived input, input is more likely to be comprehended if the learner has more relevant prior linguistic knowledge. Comprehended input, however, is different from intake. Comprehended input could be comprehended just for a moment. Intake, on the other hand, involves an attempt to integrate the linguistic information into the learner’s developing L2 system (Gass & Selinker, 2008). In this way, intake is similar to Schmidt’s (1990) notion of awareness at the level of understanding. Intake is facilitated by the quality of the analysis at the comprehended input stage, knowledge of the L1 and the L2, and if the feature is an aspect of the L1 or an innate aspect of languages. Intake involves psycholinguistic processing in which the learner compares prior knowledge, such as grammatical knowledge in the L1 and L2, to the input in order to interpret it and form, test, reject, modify, or confirm a hypothesis on linguistic aspects of the L2. A mismatch between the input and prior knowledge, for example, may lead to the learner modifying his original hypothesis to incorporate the new data, or rejecting it entirely. The next step, integration, is the use of input to modify or strengthen a learner’s internal grammar, or the storage of input for later use. Output is what the learner produces and is largely based on the learner’s developing grammar, although Gass and Selinker (2008) caution that the two are not identical, due to the potential role of personality and task factors. In Figure 14.1 (p. 481), which theorizes on the stages involved in second language
acquisition, Gass and Selinker (2008) assert that the L2 learning process does not end with output, but rather, that the model of the L2 learning process can be seen as circular, with output having an influence on how apperceived input is processed.

According to Gass and Selinker (2008), comprehended input may consist of understanding the form or meaning of a sentence, or both. A number of theories and empirical studies have explored the nature of processing for meaning and/or form (Greenslade, Bouden, & Sanz, 1999; Leow, Hsieh, & Moreno, 2008; Peters, 1985; Sharwood Smith, 1986; Hulstijn, 1989; Morgan-Short, Heil, Botero-Moriarty, Ebert, 2012; VanPatten, 1984, 1990, 2004; Wong, 2001).

As outlined by VanPatten (1996), Peters (1985) and Sharwood Smith (1986) were two early theories that emphasized the importance of processing for meaning in language acquisition. For first language acquisition, Peters (1985) theorizes that meaning is a particularly “salient feature” of input (p. 1034), and that children pay close attention to meaning and context. Peters (1985) also theorizes that second language naturalistic learners follow “basically the same” principles as she outlined for children, implying that processing input for meaning likely plays a primary role for the second language learner as well (p. 1054). In addition, Sharwood Smith (1986) argues that a learner who wishes to “succeed in a given interchange of messages” will focus on meaning (p. 243). Given the widely accepted and previously discussed assumption that L2 learner’s attention is limited (e.g., Robinson, 1995; Schmidt, 2001), it is possible that this primary focus on meaning could negatively affect form learning (VanPatten, 2004).

VanPatten (2004) theoretically explores the limits to human processing. Specifically, VanPatten (2004) proposes a series of input processing principals based on three main claims in the SLA literature (pp. 56-57):

“(1) Learners are focused primarily on the extraction of meaning (Fæerch & Kasper, 1986; Krashen, 1982).
(2) Learners must somehow ‘notice’ things in the input for acquisition to happen (Schmidt, 1990)
(3) Noticing is constrained by working memory limitations regarding the amount of time they can hold and process during on line (or real time) computation of sentences during comprehension (e.g., Just & Carpenter, 1992).”

VanPatten’s Primacy of Meaning Principle

VanPatten’s (2004) first principle, known as the Primacy of Meaning Principle, holds that input is first processed for meaning, and second for form. VanPatten (2004) explains the implications of this principle by noting that the learner, in search of meaning, and given his limited cognitive resources, “tend to” or are “more likely” to process those aspects of the input that allow them easy access to meaning. According to VanPatten (2004), this implies that the learner “processes content words before anything else” (p. 8), “lexical items as opposed to grammatical forms” (p. 9), and “nonredundant” and “meaningful” morphology before “redundant” and “nonmeaningful” morphology (p. 11). As processing resources are assumed to be constrained by working memory, the Primacy of Meaning Principle implies that processing for form may be limited in situations where the learner expends too many of their processing resources for meaning through lexical items and content words. This is most likely to occur in the case of “beginning” and “intermediate” learners who have yet to automatize their processing of meaning (VanPatten, 1984; VanPatten, 2004, p. 9), and in the case that the forms in question either do not convey meaning, or convey meaning that is already expressed by content words. In the same way, the principle suggests that when a learner is given tasks that require her to expend high levels of processing resources to process for form and meaning at the same time, without
allowing her any additional time to compensate for the added burden of processing for form, her comprehension will be negatively affected.

Empirical Research

Hulstijn and Hulstijn (1984) and Hulstijn (1989) were the first studies to explore the effect of focusing attention on form or content, and, in the latter study, the effect of focusing attention on both. Hulstijn and Hulstijn (1984) empirically examined the effect of time pressure and focus of attention on meaning or form. In their study, 32 adult learners of Dutch each heard and retold stories under four conditions: (1) Information/Fast; (2) Information/Slow; (3) Grammar/Fast; and (4) Grammar/Slow. The Information conditions gave participants specific instructions to focus on meaning, and were told that they would be scored for meaning, while the Grammar conditions gave participants specific instructions to focus on grammar, and told that they would be scored for grammar. Additionally, in the Fast conditions, participants were told to retell the stories quickly, while in the Slow conditions, participants were told to take their time. Interestingly, there was no difference in score between the Fast and Slow conditions. This was not expected, given that, with more time, learners should be able to take advantage of additional cognitive resources. It may be, given that there was no specific time limit, that time was not constrained enough in order to have an impact on the results. Also, Hulstijn and Hulstijn (1984) found that participants performed better when they were told to focus on meaning in terms of the content of their retellings, although this finding was not significant. Hulstijn and Hulstijn (1984) did, however, find significant results with regards to the focus on grammar condition. Specifically, when participants were told to focus on grammar, they performed significantly better in terms of the two grammatical items that the authors measured: (1) inversion of the
subject and the finite verb; (2) placement of the finite verb in final position. This finding supports the limited-capacity model of attention (e.g., Robinson, 1995; Schmidt, 2001), and suggests that focusing on form may be able to be disassociated from focusing on meaning.

Nonetheless, three issues limit the strength and the scope of the study’s findings. First, given that no group focused on both Form and Meaning, it is impossible to disassociate the effect of focusing on one aspect (Form or Meaning) from ignoring the other. Second, the study did not ensure that participants were focusing on grammar or meaning when instructed to do so. This is especially problematic given that each participant read 68 separate stories, each with one of the four different instructions, which could have potentially been confusing to participants. Third, even in the event that they were able to verify that the participants followed instructions faithfully, this would have led learners to focus on meaning or grammar not only in the input phase, but, also in the testing phase, given that they were told that they would be graded on either meaning or form. Therefore, as Chaudron (1985) points out, the effect of focusing on meaning or form during the input processing phase cannot be disassociated from the effect of focusing on meaning or form during the testing phase.

Hulstijn (1989) examined the impact of focus on attention on learner comprehension and production in two separate experiments. In the first of the two experiments, 123 intermediate Dutch learners were pseudo-randomly assigned to four conditions: Form Group, Meaning Group, second Meaning Group, and Form and Meaning Group. The target structure was word order in Dutch interrogative complex sentences. In the Form Group, participants were given pieces of Dutch interrogative complex sentences in eight fragments, and had the task of putting the fragments together to form a grammatically correct question. In the Meaning Group, participants were asked to think about these questions (the same interrogative complex sentences) and answer
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

them by writing down their opinion. The second Meaning Group received the same task as the Meaning Group, except that the questions disappeared after 10 seconds. In the Form and Meaning Group, participants were simply shown the questions and asked to pay attention to form and meaning without any additional task. In the second experiment, 80 native Dutch speakers were placed into similar conditions, except that an artificial surface structure was superimposed on Dutch content words, and only one Meaning Group, similar to the second Meaning Group in the first experiment, was used. Subsequent to both experiments, participants took the Cued Recall test, where participants were asked to recall the nine target questions that were previously presented to them based on an adverbial phrase/temporal cue. On the Cued Recall test, the Form group performed significantly better than the Meaning group(s) on production of the grammatical structure, while the two Meaning groups scored significantly higher on comprehension of the content. The results of Hulstijn (1989), therefore, appear to suggest that both that a focus on form may be able to be disassociated from a focus on meaning, and vice versa, and in terms of both comprehension and production. Additionally, as Hulstijn (1989) notes, the study demonstrates that attention to form is sufficient for production to occur, although, given that the Meaning group also improved from the pre-test to the post-test, Hulstijn (1989) does not establish that attention to form is necessary. An interesting finding was that participants in the Form and Meaning condition performed significantly better than participants in the Form condition on production of the grammatical structure. In the first experiment, it was also found that the Form and Meaning condition performed significantly better than the Meaning condition on comprehension. These results appear to suggest that participants are not only able to pay attention to form and meaning without having any detriment to performance, but actually benefit in terms of form, and even potentially meaning, when compared to only focusing on one
or the other, providing possibly contradicting evidence to the well-adopted supposition that attention is limited in its capacity (e.g., McLaughlin, 1987; Robinson, 1995; Schmidt, 2001, VanPatten, 2004). Hulstijn’s (1989) Form and Meaning results, however, must be interpreted cautiously, as participants in this condition did not perform any specific task to ensure that they paid attention to both, or even one, of these two aspects of the sentences. Moreover, the Form and Meaning group was not pressured by any specific time limit, potentially allowing participants to compensate for the additional burden of paying attention to form by allowing themselves more time to complete the task.

In VanPatten’s (1990) study of simultaneous attention to form and meaning, participants in all conditions had the same amount of time to process the input and participants in the form and meaning condition were given a task to ensure that they attended to form. In the study, 202 first-semester, fourth-semester, and third-year Spanish participants first listened to a warm-up passage. They then listened to a passage on Latin American inflation, under four conditions: (I) attention to content/meaning only; (II) attention to content/meaning and noting the lexical item *inflación*; (III) attention to content/meaning and noting the definite article *la*; (IV) attention to content/meaning and noting the verb morpheme –n. VanPatten (1990) operationalized attention to content by instructing participants to listen for meaning and telling them that their comprehension of the passage would be assessed afterwards. VanPatten (1990) operationalized attention to form by asking participants to place a check mark on a piece of paper as they heard each instance of the item they were attending to, and eliminating participants that did not do so in at least 60% of instances. Comprehension was assessed by a free written recall in English. VanPatten (1990) found that participants in the attention to content only condition (I) did not perform significantly better in aural comprehension score than participants in the attention to
both content and the lexical item *inflación* condition (II). VanPatten (1990) also found that participants in both these conditions (I, II) scored significantly higher in aural comprehension than participants in both the attention to content plus the definite article *la* condition (III) and the attention to content plus the verb morpheme –*n* condition (IV). VanPatten (1990) concluded that the negative impact on aural comprehension in two of the attention to form conditions (III, IV) suggests that attention to form and attention to meaning compete for the same processing resources.

Greenslade et al. (1999), in their conceptual replication study, expanded on VanPatten (1990) by exploring simultaneous attention to form and meaning in the written mode. Sixty-two third-semester Spanish participants from intact classes participated in the study. The procedure closely followed VanPatten (1990), except that participants read the Latin American passage on inflation, instead of listening to it. Following the results of a pilot test, participants were given 2 minutes and 30 seconds to do the reading, in order to be able to compare the results with VanPatten (1990). Similar to VanPatten (1990), participants in Groups II-IV were required to circle, underline or check each instance of the target item, and were eliminated if they did not so for at least 60% of instances. Greenslade et al. (1999) reported similar results to VanPatten (1990), with significant differences between Group I (control) and Groups III (*la*) and IV (*-n*), as well as between Group II (*inflación*) and III. Greenslade et al. (1999) differed from VanPatten’s (1990) results in that it did not find a significant difference between Group II and IV. Also, Greenslade et al. (1999) reported a significant difference between Groups III and IV, not found in VanPatten (1990). Greenslade et al. (1999) asserted that differences between their study and VanPatten (1990) suggest that modality plays an important role in processing. Overall,
Greenslade et al. (1999) claimed that the data suggests that focusing on form interferes with comprehension in the written mode, as VanPatten (1990) found it to be in the aural mode.

Given that Greenslade et al. (1999) compared the written and aural modalities with different pools of participants, Wong (2001) empirically examined the effect of aural and written modality on simultaneously attending to form and meaning with the same pool of participants. In her study, 79 low-intermediate French students of English were assigned to either a reading modality condition (45) or a listening modality condition (34). For each modality, there were three conditions to which participants were assigned: (I) attention to content only; (II) attention to content and the content word inflation; (III) attention to content and the definite article the. The procedure was the same as in VanPatten (1990), except that for the reading modality conditions, participants in the attention to form conditions were required to underline the forms. With regards to aural modality, Wong (2001) found similar results to VanPatten (1990), with the Content only and the Content plus attention to the content word conditions performing better than the Content plus attention to the definite article condition. In the written mode, while no significant differences were found between the three conditions, contrary to Greenslade et al. (1999), the Content only condition did score higher than both of the two remaining Content plus attention form conditions. The finding of significant results in Greenslade et al. (1999), but not Wong (2001), despite both showing higher scores for the Content only condition, may have been due to the inclusion of much fewer participants in the latter (62 vs. 34), an important factor in determining the likelihood of significant findings given a difference in population especially if the difference in population is not large. Additionally, a main effect was found for Mode, and the interaction between Mode and Task approached significance ($p < .0552$). Wong (2001) concluded that the aural mode appears to constrain the learners’ limited attentional capacity more.
so than does the written mode. This would be consistent with the above argument, despite a relatively similar number of participants in the two experiments, given that the aural modality may involve a higher overall difference in the population, making it more likely that a relatively small sample size still leads to significant results. It should be noted, however, that this does not necessarily mean that learner’s attentional capacity is unlimited in the written modality. Rather, it is possible that higher demands on processing for form may be necessary in order to have a detrimental effect on participants’ comprehension.

Overall, the results of VanPatten (1990), Greenslade et al. (1999), and Wong (2001) provide support for VanPatten’s (2004) Primacy of Meaning Principle. Each of these experiments found at least one attention to form condition to have a significant negative impact on performance, with the exception of Wong’s (2001) results for the written modality, which, crucially, given the role of sample size in determining Type II error, included fewer participants than any of the other experiments including Wong’s (2001) aural-only experiment. Indeed, in all three studies, including Wong’s (2001) written-modality experiment, the average comprehension score for the attention to meaning only condition was higher than all of the remaining attention to meaning and form conditions. The results of the three studies suggest, as VanPatten (1990) originally noted with regards to the results in his study, that attention to form and attention to meaning compete for the same cognitive resources, and, when the cognitive resources are sufficiently limited, such as through a time limit as done in these three studies, attention to form can have a detrimental effect on participants’ ability to pay attention to meaning. Nevertheless, as Leow et al. (2008) point out, meaning was not adequately operationalized. In addition, among other methodological concerns, VanPatten (1990) did not determine the minimum amount of attention to form that was necessary to have an impact on meaning. It may be that participants
were only impacted by processing for form when they paid attention to many more forms than the minimum 60% of forms that VanPatten (1990) prescribed. As paying attention does not necessarily imply an effort to take in the input, and to avoid terminological confusion, the term *processing for* as opposed to *attention to* will be used henceforth to refer to the use of cognitive resources to take in input.

Leow, Hsieh, and Moreno (2008) highlighted a series of issues in VanPatten’s (1990) original research design on simultaneously processing for form and meaning, followed closely in the subsequent replication studies previously reviewed (Greenslade et al., 1999; Wong, 2001), and improved upon the design in their experiment. Some methodological issues that Leow et al. (2008) identified include the use of a polysyllabic lexical item (*inflación*) that does not allow a direct comparison to the monosyllabic morphological items (*la, -n*), the use of written recalls as the comprehension assessment task, which only represented 33% of the idea units, and the lack of an even distribution of the targeted forms in the input. The most crucial methodological issue of previous studies identified by Leow et al. (2008) is that attention to, or more specifically, processing for meaning was not adequately operationalized, as previous studies simply gave participants instructions to pay attention to meaning and told them that they would be tested on their comprehension, but did not follow-up on the participants to ensure that they did in fact pay attention to meaning. Leow et al (2008) pointed out that several participants in VanPatten (1990) commented that they forgot to process for meaning (e.g., “I don’t know, I forgot to pay attention to the meaning of the passage,” p. 295).

In their study, Leow et al. (2008) explored the effect of simultaneously processing for meaning and form in the written modality among 72 second-semester, college-level Spanish students. As in the previous studies (Greenslade et al., 1999; VanPatten, 1990; Wong, 2001),
Leow et al. (2008) assigned participants to a processing for content only condition and a series of conditions involving processing for content and processing a specific form. Unlike previous studies, however, Leow et al. (2008) used a monosyllabic lexical item (sol 'sun') and a form carrying both meaning and grammatical function (the masculine singular direct object lo), in addition to the article (la) and the morpheme (-n) used in previous studies (Greenslade et al., 1999; VanPatten, 1990;). Leow et al. (2008) were careful to distribute target forms in equal proportion to the content corresponding to the subsequent comprehension questions. They operationalized processing for form by counting forms circled and/or mentioned in the Think-Alouds. Processing for meaning, unlike in previous studies (Greenslade, 1999; VanPatten, 1990; Wong, 2001), was operationalized via a multiple-choice task and those participants who did not demonstrate a sustained effort to read the text for meaning were eliminated from the final pool. Leow et al. (2008) also discarded participants who went back to the text to look for answers. Leow et al. (2008) did not find any significant effect for condition, meaning that attention to or processing target forms, as operationalized here, did not affect comprehension. Morgan-Short et al. (2012), who employed a similar design, but added a control group for reactivity, also did not find any significant difference between groups that processed for form and meaning and those that processed for meaning only. Nonetheless, Leow et al. (2008) relied on written instructions alone to ensure that participants proceeded as quickly as possible, while Morgan-Short et al. (2012) allowed participants to proceed “at their own pace” (p. 671). This may have potentially allowed participants in the processing for form and meaning group to access more cognitive resources to compensate for the additional task of processing for form (Leow et al., 2008 and Morgan-Short et al., 2012 are discussed in more detail below under Empirical Evidence & Depth of Processing in the L2 and Methodological Issues of Empirical Research).
Depth of Processing

One of the earlier references to depths of processing was Craik and Lockhart (1972), who theorized that learners attempted to take in the input with different levels of processing that vary as to their depth. According to Craik and Lockhart (1972), depth is related to the complexity of the analysis of the input, with “greater ‘depth’ impl[y]ing a greater degree of semantic or cognitive analysis” (p. 675). They theorize that low levels of processing involve “physical or sensory features [like] lines, angles, brightness, pitch, and loudness,” while higher levels of processing are used when connections, such as “pattern recognition and the extraction of meaning,” are made between the input and past learning (p. 675).

Craik and Lockhart (1972) theorize that when input is processed at a deeper level it is more likely to be remembered. They postulate that a trace is created when an input is attended to, and that the intensity of this trace depends on the depth with which the input is processed, “with deeper levels of analysis associated with more elaborate, longer lasting, and stronger traces” (p. 675). A stronger trace is predicted not only to result in a greater ability to recognize and recall the input, but also the durability of ability across the passage of time and in the instance of subsequent diversions of attention.

Craik and Lockhart (1972) argue that learners are more easily able to process familiar input deeply than less familiar input. The implication of this argument is that time alone cannot reliably be used as a measure of depth of processing across different tasks, as less familiar tasks may take longer to process at the same level of processing as their more familiar counterparts. Therefore, according to Craik and Lockhart (1972), “amount of attention… its compatibility of
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

the analyzing structures, and the processing time available, will determine the depth to which it is processed” (p. 676).

The first major pillar to Craik and Lockhart’s (1972) theory is the finding of a number of studies that semantic processing is superior in terms of memory performance to word-structure processing under incidental conditions (e.g., Hyde & Jenkins, 1969; Johnston & Jenkins, 1971; Tresselt & Mayzner, 1960). In these studies, ranking a word as to its pleasantness, a form of semantic processing, was found to lead to significantly higher retention than checking for the letter e or estimating the total number of letters in a word (Hyde & Jenkins, 1969), writing down an adjective or noun that could be modified or modified by the word was found to lead to higher retention than writing a rhyming word (Johnston & Jenkins, 1971), and judging a word to belong or not belong in a category resulted in significantly higher retention than crossing out the vowels in words (Tresselt & Mayzner, 1960). Therefore, the finding of increased retention under the semantic processing condition is consistent regardless of the specific type of semantic and word-structure processing tasks. As previously noted, Craik and Lockhart (1972) assert that semantic processing is a higher level of processing than structural processing. Given this categorization of semantic processing, therefore, these studies suggest that higher processing leads to greater retention.

The second major pillar of Craik and Lockhart’s (1972) theory is that increased repetition within a low level of processing was shown in studies to not improve performance (Tulving, 1966; Turvey, 1967). Tulving (1966) and Turvey (1967) found that recall for novel items, both for words (Tulving, 1966) and numbers (Turvey, 1967) did not differ from recall for previously presented items. The lack of an effect of repetition within a low level of processing, combined with the finding that processing at a higher level leads to increased retention, provides support
for Craik and Lockhart’s (1972) theory that processing levels is the key factor in the development of a durable trace. Several criticisms of Craik and Lockhart’s (1972) depth of processing theory, however, have since been pointed out by Eysenck (1978) and Baddeley (1978).

Baddeley (1978), oppugning the continuous nature of the processing model, notes that Craik and Lockhart (1972) pointed out no study showing evidence of levels of processing within any one domain. He also cites a more recent study, Craik and Tulving (1975), explored in depth *in Empirical Evidence & Depth of Processing in the L1*), in which no difference was found between their two semantic processing levels or between their two phonemic processing levels.

Eysenck (1978), in his critique of Craik and Lockhart (1972), additionally pointed out that Craik and Lockhart (1972) focused on input to such an extent that they largely ignored the potential role of the retention-test. Eysenck (1978) argues that both recall and recognition tests should be used to measure participants’ memory, so that any potential test-specific effect is minimized.

Most importantly, Baddeley (1978) and Eysenck (1978) both assert that Craik and Lockhart’s (1972) theory that higher levels of processing lead to greater retention relies on circular logic in their interpretation of semantic processing studies. Eysenck argues that, because Craik and Lockhart (1972) were able to examine findings that semantic processing lead to superior retention compared to structural processing prior to theorizing which type of processing was higher than the other, they may have classified semantic processing as a higher level of processing based on those results. Eysenck (1978), therefore, argues that an independent index of depth is needed in order for the theory to be empirically tested.
Lockhart and Craik (1990) find Baddeley’s (1978) criticism that different processing levels have not been found within domains to be “reasonable” (p. 96), noting that they have since argued (Lockhart, Craik, & Jacoby, 1976, as cited in Lockhart & Craik, 1990) that processing levels might refer to “qualitatively different domains,” while “elaboration” might be a better term for the amount of analysis within a particular domain (p. 96).

Lockhart and Craik (1978) also conceded that the retrieval cue has an impact on retention (e.g., Tulving & Osler, 1968; Fisher & Craik, 1977) and that this is not entirely consistent with the original conception of the theory that the trace is dependent on the level of (input) processing. And while Tulving and Osler (1968) did find that the effectiveness of the recall cue was dependent on that recall cue present at the input stage, Fisher and Craik (1977), who did not include the recall cue at the input stage, nonetheless found recall cue to have an impact. More specifically, Fisher and Craik (1977) found that recall was highest when the retrieval cue was of the same type as the processing condition. Fisher and Craik (1977) also found, however, when controlling for retrieval cue, that semantic processing was still significantly more effective than phonemic processing. This indicates that while retrieval cue does play a role, it does not account for general differences in retention between the two processing levels.

Lockhart and Craik (1978, 1990) also admitted that “circularity is inherent… in the levels of processing approach,” (1978, p. 172). They agree with Eysenck (1978) that an “independent index of processing” is necessary, and note that “processing effort or time” might be a potential component, but argue that it is only with empirical research that a reliable independent index will be revealed (Lockhart & Craik, 1990, p. 92).

While conceding these limitations, however, Lockhart and Craik (1978, 1990) note that the theory, as a whole, has been productive. Craik (2002), in retrospect, notes that he has “seen
no evidence against the proposition[...] that more meaningful processing is usually associated with higher levels of recollection” (p. 316). Quite the contrary, research in the L1 from Craik and Tulving (1975) to Gardiner, Brandt, Vargha-Khadem, Baddeley, and Mishlin (2006) has found evidence to be consistently in favor of this most basic principle of Craik and Lockhart (1972).

Empirical Evidence & Depth of Processing in the L1

In ten separate, but related, experiments, Craik and Tulving (1975) explored the relationship between depth of processing and memory performance. The general method involved in the ten experiments was as follows: (1) Participants were exposed, either visually or verbally, to a series of questions, saw a word in a tachistoscope, and answered the question with a “yes” or a “no” by pressing one of two keys; (2) Participants took a retention test (recognition, cued recall or recall) probing their memory of the words to which they were previously exposed.

In the first experiment, participants were given five types of yes-or-no questions which varied as to their level of complexity. The first three questions asked about the structure of the word. The simplest of these questions simply asked whether or not a word was present, while the most complex of the three required the participant to determine if the word rhymed with another word that was provided. The last two questions required the participants to consider the semantic meaning of the word, and, therefore, following Craik and Lockhart’s (1972) theory, were the most complex. The participants then were given a recognition test with 40 distracters and 40 of the words presented to them and were asked to identify those words that they had previously seen. The authors found that the deeper semantic questions resulted in significantly longer latency times and higher rates of recognition, providing support to Craik and Lockhart’s (1972) theory that input that is processed more deeply is more likely to be retained. Interestingly, while
response type (yes vs. no) was a significant factor for recognition, with yes answers leading to greater recognition than no answers, it was not significant for latency. Nonetheless, the authors noted ceiling effects in the scores obtained.

Experiment 2 was similar to Experiment 1 except that it reduced the number of question types and included additional distracters in the recognition test in order to control for ceiling effects. The experiment included the two structure question types described in the previously paragraph, and one semantic question type. In addition, the recognition test included 120 distracters. As in Experiment 1, semantic questions resulted in longer latency times and higher recognition scores when compared to the structure question types, while response type was only a significant factor for retention.

Experiment 3 explored the relationship between depth of processing and recall. The same three question types in Experiment 2 were used in this experiment. In addition, in order to increase recall, fifty percent of the words appeared twice. Following the questions, participants were given a one-minute break and subsequently took a recall test that required participants to write down all the words they were able to. The semantic questions resulted in higher latency times and higher retention, extending the relationship between depth of processing and memory established in the two previous experiments to include written production. Also, as in the previous two experiments, response type was significant for retention, but not for latency.

The only difference in experiment 4 was that participants were told prior to seeing the questions that they would be subsequently asked to recall the words they were shown. The same significant findings mentioned in the previous paragraph for experiment 3 were found in experiment 4.
Craik and Tulving (1975) argue that the results of experiments 1 through 4 provide support for the general concepts of Craik and Lockhart’s (1972) theory, but also assert that the results suggest a need for modifications to this theory. Generally, they observe that the results show that deeper processing, operationalized as semantic and not structural questions, lead to greater retention measured in terms of both recognition and recall ability. Nonetheless, they contend that the finding that response type was significant for retention, but not for latency, contradicts their postulation that more time leads to deeper processing. Indeed, according to Craik and Tulving (1975), the finding is indicative of a disassociation between time and depth of processing, as the materials were the same for both response types, and, therefore, if time were a crucial factor in depth of processing, it would have been expected, they argue, that time and retention showed the same pattern of significance. An alternative explanation, however, might be that questions with a ‘yes’ answer as opposed to a ‘no’ answer are fundamentally different. Indeed, Craik and Tulving (1975) themselves later assert that ‘yes’ responses involve more integration than ‘no’ responses (see below), possibly explaining the time differential. Craik and Tulving (1975) also argue that the finding that question type was a significant factor under the intentional learning condition in Experiment 4 contradicts their original postulation that semantic analysis as opposed to structural analysis (i.e., deep vs. shallow processing) is the key factor leading to higher retention. The reasoning is that participants, made aware of the subsequent retention test in this experiment, would “surely analyze and perceive the name and the meaning of the target word with all three types of questions [anyway]” (p. 279), and, thus, given that a semantic analysis is present for all question types, it would be expected that question type not be a significant factor in retention if semantic analysis is the key factor in retention differences. Nonetheless, the assumption that all participants will “surely” semantically analyze the words
just because they are told they will be tested on them does not appear to be well founded. Certainly, it is possible that many participants would indeed semantically process the words to prepare themselves for the retention tests, but it also seems likely that some would not. Given that the semantic question type guarantees a semantic analysis of the word, as such an analysis is necessary to answer the question, and the intentional nature of the learning design only makes it likely that most participants will do so, the difference in retention with regards to question type seems to still be explained by a theory which posits semantic analysis as key to retention.

Experiment 5 was designed to test Craik and Lockhart’s (1972) postulation that time is not a reliable measurement of depth of processing across different task types. In the experiment, two tasks were given to participants, a time-consuming, but structure-related (low depth of processing/non-semantic) task and a quick, but semantic-related task. For the structure-related task participants determined the vowel-consonant composition of the word in question, while, for the semantic-related task, participants decided whether or not a series of words could fill in the blanks of the sentences presented. The results of the experiment showed that the structure-related questions resulted in longer latencies, but lower retention, providing support for Craik and Lockhart’s (1972) postulation that time is not reliable in measuring depth of processing.

Craik and Tulving (1975) also explore the reasoning behind the significance for response type (yes/no). They hypothesize that questions with a ‘yes’ answer might have allowed for greater integration between the word and the question, and that this integration may have been facilitative to participants’ memory of the word. For example, they note, there is intuitively a greater integration between the question “A four-footed animal?” and “BEAR,” which is a yes-response question-word pair, and “A four-footed animal?” and “CLOUD,” which is a no-response question-word pair (pp. 281-282).
Experiment 6 was designed to test this hypothesis. In this experiment, the questions required the participants to relate the size, depth, width, height, weight, temperature, sharpness and value of the word and that of another entity contained in the question. For example, one question-word pair was “Sharper than a fork” and “Needle,” and required a ‘yes’ response, while another was “Sharper than a fork?” and “Club,” and required a ‘no’ response (p. 282). Thus, the questions and their respective words allowed for equal integration for both yes and no response types, and, if Craik and Tulving’s (1975) hypothesis is correct regarding the reasoning for the difference in findings between ‘yes’ and ‘no’ response types, the expectation would be, in this experiment, where integration is equal with both, that no such difference would be found. This was indeed the case, with no significant difference being found between the two response types in the subsequent recall test.

Experiment 7 explored the impact of encoding elaboration on retention. In the experiment, three different types of semantic sentences with blanks were used, which varied according to complexity: (1) simple; (2) medium; (3) complex. Participants were presented with a word for each sentence, and were required to determine whether or not the word made sense in the context. Following this, as in Experiments 3, 4 and 6, participants took a recall test. Craik and Tulving (1975) found that for ‘yes,’ but not ‘no,’ response-type question-word pairs more complex sentences lead to significantly higher recall. Given that each question type level involved semantic processing, the results show that additional variables, such as, in this case, complexity or “encoding elaboration,” are related to memory performance. They fall short, however, of providing evidence for Craik and Tulving’s (1975) contention that “degrees of elaboration” (p. 279) are a better predictor of memory than depth of processing, given that the experiment did not allow for a comparison between semantic vs. non-semantic question types.
Future research should consider including question types that vary by both their semanticity and their complexity.

Experiment 8 explored the impact of uniqueness in question type on the retention rate. Three groups of participants were created that received a different number of each question type. Group 1 was required to answer many category question-word pairs (40), several rhyme question-word pairs (16), and a few case question-word pairs (4), while group 2 answered 4, 40, and 16, respectively, and group 3 answered 16, 4, and 40 respectively. The reasoning is that if uniqueness is a factor, the expectation would be that those that answered fewer of a particular question-word type would be able to recall the words corresponding to this question-type than words of question-types they answered more frequently. Craik and Tulving (1975) found that while set size did not affect overall retention, it did have an effect on retention of words of rhyme question-word pairs, indicating that uniqueness may be a factor on memory only under certain conditions.

Experiment 9 tested the impact of question-type in a classroom setting. Sixty words were presented on a projection screen, and different students were asked different questions about the words. Each student answered, on paper, a total of six questions per word that varied by question type (case, rhyme, category) and response type (yes/no). Participants then took a recognition test with the 60 words and 120 distracters, and were asked to select those words to which they had been previously exposed. Craik and Tulving (1975) found that the results in this experiment performed in a classroom paralleled results from previous experiments, in which words of category (semantic) question-word pairs were better retained than words from rhyme and case questions. Experiment 9, therefore, expands the scope of the Craik and Tulving’s (1975) findings to include natural settings, or, more specifically, to the classroom. Craik and Tulving (1975) also
argue that the experiment, which did not take time into consideration, underscores the importance of depth of encoding, as the author refers to the concept here, as opposed to time as a viable predictor of retention. Nonetheless, while the connection between depth of encoding and retention is indeed supported by this study, it is unclear how this particular experiment negates the impact of time since it was not measured in this study.

Experiment 10 examined the role of motivation in retention. The experiment was similar to Experiment 9, except that three groups were created, which varied based on the type of words, those corresponding to case, rhyme, or category questions, whose retention was better financially compensated. Financial compensation was not found to have an effect on retention, and thus, no role for motivation can be ascribed based on the findings of this experiment.

In summary, experiments 1 through 4 established that question-type and latency have an effect on retention. While response-type was a significant factor for question-type, however, it was not a significant factor for latency. Craik and Tulving (1975) argue that this suggests that time is not an important factor in depth of processing. It was suggested here, however, as noted by Craik and Tulving (1975) with regards to experiment 6, that different response types change the level of integrativeness of the question-word pair, potentially explaining why response-type was not a significant factor for time. Experiment 7 find support for Craik and Tulving’s (1975) assertion that complexity may play a role in retention, however, the authors’ claim that the experiment shows that “degrees of elaboration” (p. 279) are more reliable predictors of retention does not appear to be supported by the evidence. Experiment 8 found evidence that uniqueness may, under certain conditions, play a role in retention, while experiment 10 found no evidence that motivation plays a role. Finally, experiment 9 extends the general findings of the previous experiments that depth of processing plays a role in retention to the classroom setting.
Subsequent studies have found continued support for the importance of processing levels in vocabulary retention (e.g., Gardiner, 1988; Gardiner, Java, & Richardson-Klavehn, 1996; Gardiner, Brandt, Vargha-Khadem, Baddeley, & Mishlin, 2006; Java, Gregg, & Gardiner, 1997; Khoe, Kroll, Yonelinas, Dobbins, & Knight, 2000; Rajaram, 1993).

Gardiner (1988) empirically re-examined Craik and Lockhart’s (1972) depth of processing theory and explored the potential role of awareness in his first experiment. In the experiment, 32 undergraduates in London were randomly assigned to process words either semantically or phonemically. Under the semantic processing condition, for each word, participants were asked to write down the first word that came to mind that was similar in meaning to the given word. Under the phonemic processing condition, participants were asked to write down the first word that came to mind that rhymed with or sounded similar to each target word. In the recognition test, the 36 target words were combined with 36 lure words in a list, and participants were required to circle those words they believed were the target words they were exposed to earlier. In addition, participants were asked to identify each selection as either “remember” or “know,” indicating, respectively, that their selection was due to remembering the experience they had with the word in the processing stage, or that they believed the word was presented to them earlier but did not remember the processing experience. Participants were explicitly discouraged from guessing. Gardiner (1988) found evidence of a processing effect, but only for the “remember” response-type, suggesting that the advantage of deeper, or, more specifically, semantic processing, is due to its superior explicit memory of items, or, as Gardiner (1988) suggests, following Tulving’s (1983, 1985) schema of memory systems, improved episodic memory, as opposed to semantic memory. Episodic memory can be disguised from semantic memory in that the former involves “autonoetic consciousness” or “an awareness of the
event,” whereas the later refers to general “symbolic” knowledge about the world (Tulving, 1985, p. 3).

Rajaram’s (1993) experiment 1 conceptually replicated Gardiner (1988), adding modality as an additional variable. The experiment was otherwise similar to Gardiner (1988), except that it employed a within-subjects design, in which all 16 participants participated in four blocks of processing words which included all possible combinations of low (phonemic) and high (semantic) processing levels with auditory and visual modalities, and employed a 1-hour, as opposed to 10 minute, retention interval. Rajaram (1993) also did not explicitly discourage participants from guessing. Rajaram (1993) did not find any evidence that modality played a role in the results. His results were similar to those of Gardiner (1988), except that for the “know” response type, phonemic processing was found to be superior to semantic processing in terms of scores on the recognition test. This compares to Gardiner’s (1988) result of finding of no significant difference between the two processing levels. Rajaram (1993) contend that the finding of disproportionate number of remember responses for semantic processing compared to phonemic processing indicates that awareness is not the only factor involved in recognition memory. Rajaram (1993) argues, on the basis of these results, as well as the results of previous studies (e.g., Gardiner, 1988), that remember and know response-types reflect “two bases of recognition memory” (p. 99), with remember responses reflecting episodic memory and know responses reflecting semantic memory. Rajaram (1993) clarifies, however, noting that “know” responses, given that the participant is instructed to reflect on the source of his memory, may not be entirely implicit, and that remember and know responses reflect a continuum and not a dichotomy between explicitness and implicitness. The results of Rajaram (1993) provide partial support to Craik and Lockhart’s (1972) theory, in that they show overall gains for semantic
processing, although the increase in know responses for the phonemic processing condition points to a potential limitation in Craik and Lockhart’s (1972) theory. Gardiner, Java, and Richardson-Klavehn (1996) point out a methodological concern, however, with regards to conclusions drawn on the basis of the know responses in Rajaram (1993). They note that given that in this experiment participants were not explicitly discouraged from guessing, that know responses may have been used as a proxy for guessing, a possibility that is highlighted by the relatively high percentage of false alarms for the know response-type. Gardiner, Java, and Richardson-Klavehn (1996) addressed this potential methodological concern.

Gardiner et al. (1996), in two experiments, re-examined the potential influence of awareness on retention under semantic and graphemic processing conditions with the addition of a “guess” response option and a “stronger manipulation” of the processing levels (p. 116). In the first experiment, which followed a within-subjects design, 16 undergraduates processed 20 words under the semantic condition and 20 words under the graphemic condition. Under the semantic condition, participants were required to write down a word related in terms of meaning, while, under the graphemic task, participants were asked to write down two letters not present in each word. After a day interval, participants took a recognition test, similar to those used in Gardiner (1988) and Rajaram (1993), except that participants also had the option of choosing “guess” as the reasoning for their selection. As in previous studies (Gardiner, 1988; Rajaram, 1993), Gardiner et al. (1996) found that, for the remember response-type, processing level was significant, with higher retention levels for the semantic processing condition. Contrary to Rajaram (1993), however, Gardiner et al. found no evidence that the opposite was true for the know response-type, finding instead that the guess response-type was significantly more present for words recognized under the graphemic processing condition. Gardiner et al. (1996) also
found that, for the guess response-type, target words were identified equally as frequently as non-target lure words. Gardiner et al. (1996) found similar results in their second experiment, which included a forced-choice recognition test and 5 to 7 day time interval. The lack of an increase in *know* responses when participants were allowed to indicate that their selection was the result of a guess suggests that semantic processing, when compared to phonemic processing, does not result in less implicit retention, but, rather, only leads to gains in explicit retention. The results of Gardiner et al. (1996) provide strong support for Craik and Lockhart’s (1972) depth of processing theory, in that they show that a higher level of processing results in overall gains without negatively affecting retention of a relatively more implicit nature (i.e., those of a *know* response type). One potential methodological concern, however, is that participants may not be differentiating “remember” and “know” responses as instructed, a possibility that was addressed in Java, Gregg, and Gardiner (1997).

In the first of two experiments, Java et al. (1997) qualitatively explored the nature of participants’ memory associated with “remember” and “know” response-types. In the experiment, 25 participants were presented with 36 words, one word at a time, and asked to commit these words to memory. Unlike in previous studies (Gardiner, 1988; Gardiner, Java, & Richardson-Klavehn, 1996; Rajaram, 1993), no particular instruction was given with regards to the method of memorization. Subsequently, participants were given a recognition test with 72 words that required them to identify the words that had already been presented to them, as well as select either “remember” or “know” as the reason for their memory of each word. Participants were then asked to “describe in detail why [they] decided a word” had already been presented to them and “why [they] decide[d] it [was]” because of a “remember” or “know” reason (p. 189). Java et al. (1997) found that 76% of the remember responses were given descriptions, compared
to 2% of know responses, and that the descriptions given coincided with the response type. For example, *remember* responses included, among other things, “images/associations,” “stories,” “personal anecdotes,” and “lists” (pp. 190-191) while the *know* descriptions largely described the participants’ degree of familiarity with the words. The results of this experiment appear to indicate that participants are able to distinguish between *know* and *remember* response-types as instructed. Khoe, Kroll, Yonelinas, Dobbins, and Knight (2000) examined the potential role of test-type.

Khoe, Kroll, Yonelinas, Dobbins, and Knight (2000) compared the recognition test used in previous studies to a forced-choice recognition test with healthy participants and amnesiacs, who are deficient in recollection, in order to determine the degree to which each test depends on recollection versus familiarity. In the first of four experiments, 32 healthy participants were required to encode 50 words at a shallow level and 50 words at a deep level. Encoding words at a shallow level involved participants typing in two vowels that were not present in each word, while for deep encoding, participants were required to type in an associate word. One week later, participants completed one of two recognition tests: (1) the yes-no recognition test; and (2) the forced-choice recognition test. The yes-no recognition test was similar to those used in previous experiments (e.g., Gardiner, 1988; Rajaram, 1993), in which participants were presented with one word at a time and asked to indicate if they saw the word before in the experiment. In the forced-choice recognition test, participants were presented with two words at a time, one of which was a previously presented word and one of which was a novel word. As in previous experiments, participants were required to indicate whether their choice was due to a general familiarity with the item, “know,” or due to a specific recollection of the encoding, “remember.”

While, as in previous experiments, a significant difference was found for encoding condition
(i.e., processing level), no such difference was found between the two test types. Similar results were found in the second experiment, which varied from the first in that the test variable was within-subjects, and in the third experiment, in which only half of the participants were required to provide a response-type.

Four amnesiacs, 6 age-matched healthy participants (62-73 years old) and 12 undergraduates took part in the fourth experiment. In this experiment, all participants processed 200 words under the shallow processing condition, and were given the standard recognition test for two-thirds of the items and the forced-choice test for the remaining third. For the amnesic patients, the words and tests were given in four separate blocks as opposed to all at once. As in the previous three experiments, no difference was found in terms of retention between the standard recognition test and the forced-choice recognition test. As the amnesiacs primarily rely on familiarity as opposed to recollection, Khoe et al. (2000) argue that the result indicates that “familiarity does not differentially support performance on [the two recognition tests]” (p. 1340).

In another experiment involving amnesiacs, Gardiner, Brant, Vargha-Khadem, Baddeley, and Mishlin (2006) again explored memory type, this time in relation to depth of processing condition as opposed to recognition test type.

Gardiner, Brandt, Vargha-Khadem, Baddeley, and Mishlin (2006) explored the type of memory elicited by semantic and phonemic processing by empirically comparing the memory test results of an amnesia patient, Jon, with impaired episodic memory, to those of 8 participants closely matched in age and IQ. In the experiment, Jon and the 8 participants were presented with 30 different words, and were required to process 15 words semantically and 15 words graphemically. Under the semantic processing condition, participants rated the words for pleasantness on a 1-3 scale, while, under the graphemic processing condition, they counted the
number of syllables in each word. They were then given a recognition test similar to that used in Gardiner et al. (1996). Gardiner et al. (2006) found that both Jon’s performance and the control group’s performance was significantly affected by processing level condition, with semantic processing leading to higher retention scores. In addition, Jon used remember as a response-type to a similar degree compared to the control group, and only selections identified with the remember response-type were significantly affected by processing condition. A significant difference was found, however, between Jon’s scores and the scores of the control group, with the control group outperforming Jon under the semantic processing condition in terms of retention. In other words, Jon did not benefit from the semantic processing condition as much as the control group. Similar to the first except that response-types were not included, and degree of imagery, instead of pleasantness, was used for the semantic processing condition, the second experiment empirically examined the impact of semantic vs. graphemic processing on retention. The results showed that while Jon and the control group were impacted by level of processing, the control groups performed significantly better in terms of retention on the semantically processed items. In addition, Jon selected non-target lure items much more frequently than the control group. Given his analogous use of the remember response-type, associated with a specific memory of the experience with the target word, and the significant impact of the processing condition on his retention, Gardner et al. (2006) suggests that the results of these experiments seem to suggest that his performance does indeed reflect episodic memory. Nonetheless, if this were indeed the case, it is not clear why Jon’s scores were not simply lower overall, to represent a lower memory overall as opposed to a deficit in episodic memory, as opposed to being lower only for those words that were semantically processed. Gardiner et al.’s (2006) third and fourth experiments shed light on this conundrum.
Gardiner et al.’s (2006) third and fourth experiments compare retention between phrases which are read and enacted and phrases which are simply read for Jon and the control group. In the third experiment, participants read and acted out 24 of 48 phrases and simply read the remaining phrases. Participants were then given a retention test similar to Gardiner et al.’s first experiment. Unlike the control group, Jon’s performance was not significantly impacted by task enactment. Moreover, the control group performed significantly better than Jon under both the enactment and read condition and the read-only condition. In the fourth experiment, which added phrases that were bizarrely read and phrases which were bizarrely enacted, Jon’s performance was similarly deficient, in addition to being inferior on the bizarrely read and enacted phrases. Moreover, Jon was unable to recall details of the enactments, unlike members of the control group. The results on experiments 3 and 4 indicate that, unlike what Gardner et al. (2006) suggested might have been inferred by the first two depth of processing experiments, that the episodic memory type is one that Jon does not likely use, or at least one that he uses to a more limited extent. Indeed, in light of the results of experiments 3 and 4, the findings of the first two experiments take on new meaning. Given that Jon does not appear to use episodic memory, his superior performance on the semantic processing tasks may indicate that semantic processing involves an integration of episodic and semantic memory, as opposed to purely being episodic, and, potentially, given Jon’s higher retention under the semantic processing condition, that the semantic memory involved in semantic processing may be superior to that of graphemic processing. In addition, Jon’s selection of the remember response-type, given his lack of episodic memory, suggests caution should be taken when relying on amnesic participants’ self-selected responses. Indeed, Gardiner et al. themselves note that “it seems more likely that Jon used ‘remember’ responses when he had strong feelings of familiarity and was highly confident of his
memory” as opposed to “genuine experiences of remembering” (p. 943). Given amnesic patients the option to label their retention as having been the result of episodic memory itself seems questionable, given the assumption that they are deficient in this memory type, and logically have an impaired vision of what exactly remembering an event implies. At the very least, future research using amnesic patients should use a triangulation of measurements to explore the episodic memory to which they do have access.

In summary, research in the L1 provides substantial evidence in favor of a role for depth of processing in retention. In all of the studies reviewed here, deeper, semantic processing lead to higher retention (Gardiner, 1988; Gardner et al., 1996, 2006; Khoe et al., 2000; Rajaram, 1993). The majority of these studies reported that this higher retention was associated only with the remember response-type, with the know response type being unaffected by processing condition (Gardiner, 1988; Gardner et al., 1996, 2006). Rajaram (1993), however, found that a lower level of processing actually led to lower retention scores for the know response type. As Gardner et al. (1996) pointed out, however, this study did not discourage participants from guessing, and participants may have used the know response type to do so. Indeed, Gardner et al. (1996), who included a guess response for participants to choose, but otherwise conducted a similar experiment to Rajaram (1993), did not find any evidence of the lower level of processing leading to lower retention scores. It has been suggested, given the relationship between remember responses and processing level, that the advantage provided by the higher processing condition is due to its enhancement of episodic memory (e.g., Gardiner, 1988; Rajaram, 1993). Nonetheless, a recent experiment in which an amnesic patient also saw enhanced retention for those words that were semantically processed, suggests that higher levels of processing may also be related to semantic memory (Gardner et al., 2006). Further substantiation of these results would provide
additional credence to Craik and Lockhart (1972) in that they would show that higher levels of processing are beneficial in a more general sense.

Diverging from the original meaning in Craik and Lockhart (1972), in which deep processing implies semantic processing and shallow processing refers to phonemic or graphemic processing, depth of processing in second language acquisition refers to the strength of the connections made within one domain. In this way, depth of processing in second language acquisition is similar to Lockhart and Tulving’s (1975) degrees of elaboration.

Empirical Evidence & Depth of Processing in the L2

As previously noted, VanPatten (2004) defines processing in SLA as making a connection between form and meaning. Depth of processing in SLA, therefore, can be seen as referring to the strength of this connection, with higher levels of processing referring to a stronger connection. An early study on depth of processing (referred to as “Attention(al) Level”) modified the input and the instructions in order to attempt to establish processing levels (Shook, 1994).

Shook (1994) empirically examined the impact of three attentional levels on participants’ intake and production of the Spanish present perfect tense and relative pronouns. In his study, 125 first and second-year students of Spanish read, on two separate days, two passages, each containing six instances of the target grammatical item. Subsequently on each of these two days, participants took a production test and a recognition test based on the grammatical item targeted in the passage (present perfect tense or relative pronouns). Participants were randomly assigned into three Attentional Conditions: (1) Control group; (2) Attention drawn to the targeted items; (3) Attention drawn to the targeted items plus focus on forming grammatical rules. Attentional
Condition 2 was operationalized by presenting participants with a passage that included the targeted grammatical items bolded and in uppercase. In addition to this modification, Attentional Condition 3 included instructions to “come up with a rule for the use of [the grammatical items]” (p. 70). Shook (1994) found that participants in Attentional Conditions 2 and 3 performed significantly better on the production and recognition tests than participants in the Control group, however no significant differences were found between Attentional Conditions 2 and 3. This study suggests that processing, but not depth of processing, plays a facilitative role in intake and production. Nonetheless, as Gass, Svecs, and Lemelin (2003) point out, manipulating input, as done in Shook (1994), only makes it “more or less likely that learners will focus attention on something” (Gass et al. 2003, p. 508), and therefore falls short of demonstrating that participants are actually paying attention to the extent that is intended by the conditions, or, potentially, even at all.

Gass et al. (2003) investigated the effect of focusing participants’ attention on syntactic, morphosyntactic and lexical items on their ability to judge the grammaticality of sentences, in the case of syntactic and morphosyntactic items, and on their ability to translate sentences for lexical items. In their study, 34 students of Italian with varied levels of proficiency read three stories, each with several instances of one type of item, in addition to a series of additional tasks. Participants were placed into one of four groups: (1) Focused attention for syntax and lexicon; (2) Focused attention for morphosyntax; (3) Focused attention for syntax; (4) Focused attention for morphosyntax and lexicon. Focused attention was operationalized by underlining the structures/words, instructions to focus on the structures/words, and questions about the structures. Additionally, in the case of focused attention to syntax and morphosyntax, participants were given a presentation of the pertinent rule, and a sentence transformation task,
while focused attention to lexicon included instructions about guessing meaning, and practice guessing meaning. Gass et al. (2003) found that focused attention had a significant impact on first-year learners’ performance for all language items, while for second-year learners a significant impact on performance was only found for lexical items. From the results of this experiment, Gass et al. (2003) conclude that focused attention has an impact on short-term learner development. A key methodological issue with Gass et al.’s (2003) study, however, lies with the multitude of variables including input enhancement, that were used to operationalize focused attention, which leads to some uncertainty regarding which variable or variables had an impact on learner development.

Laufer and Hulstijn (2001), in their theoretical approach, conceptualize a potential framework for depth of processing for vocabulary learning in the form of their Task-Induced Involvement Construct. Incorporating both a motivational and a cognitive dimension, Laufer and Hulstijn’s (2001) levels of Task-Induced Involvement in effect operationalize depth of processing for vocabulary learning by taking the sum of three components: need, search, and evaluation.

With regard to need, Laufer and Hulstijn (2001) note that instrumental motivation has been found to improve performance in empirical research (Eysenck & Eysenck, 1982; Gardner & MacIntyre, 1991), but that no study has explored how tasks might be useful in stimulating motivation. They also point out that while a concept of ‘need’ has been used in SLA literature (Crookes & Schmidt, 1991; Oxford & Shearin 1994). Crookes and Schmidt (1991), citing Keller’s (1983) general theory of motivation, argued that an aspect of motivation is the need of learners to be successful, and that learners who “think they are likely to succeed are more highly motivated” (p. 482). Oxford and Shearin (1994), in addition to noting that the desire to be
successful is an important need, point out the need to fulfill a language requirement, as well as psychological and instrumental needs as potential motivators (24). Laufer and Hulstijn (2001) note that, in the exploration of potential needs of language learners, researchers had yet to explore the potential role of specific language task needs. Laufer & Hulstijn (2001), addressing this gap, while at the same time preserving the “need for achievement” (p. 8) aspect of the definition of need, define need as the degree to which a learner is driven by both the nature of the task to obtain knowledge, or, more specifically in the context of vocabulary learning, to obtain the meaning of unknown words. Laufer & Hulstijn (2001) operationalize moderate need as a drive “imposed by an external agent” and strong need as a self-imposed drive (p. 14).

Laufer and Hulstijn (2001) were also influenced by Craik and Lockhart’s (1972) depth of processing theory. As previously noted, Craik and Lockhart’s (1972) attributed retention to processing level (1972), Craik and Tulving (1975) added that the degree of elaboration within each particular processing level also played a role. However, Craik and Lockhart (1972) and Craik and Tulving (1975) did not precisely operationalize processing levels or degrees of elaboration within each processing level. The last two components of Laufer and Hulstijn’s (2001) Task-Induced Involvement construct, search and evaluation, address this gap within the field of vocabulary learning. The search component refers to the “attempt to find the meaning” (Hulstijn and Laufer, 2001, p. 543) of a word that is unknown to the learner, and therefore can be seen as establishing at least a minimum level of processing. The evaluation component refers to the degree of analysis that the learner performs in order to ensure understanding or use the word correctly, such as by examining the various entries in a dictionary. As appropriately using the word requires the learner to not only understand the meaning of a word but also know how to use
it in a sentence, this type of evaluation is seen as being ‘high,’ while using a dictionary to understand a word in the process of reading is said to involve ‘moderate’ evaluation.

In order to calculate the total Task-Induced Involvement, Laufer and Hulstijn (2001) propose that tasks be assigned one point for each of its moderate components and two points for each of its strong components. Laufer and Hulstijn (2001) cite numerous studies in which tasks or conditions with higher Task-Induced Involvement lead to higher retention (e.g., Cho & Krashen, 1994; Hulstijn, 1992; Knight, 1994). In Cho and Krashen’s (1994) longitudinal study, it was found that those participants that used a dictionary more frequently during their leisure reading in a foreign language learned more vocabulary. Knight (1994), came to a similar finding in a controlled experiment, namely that those with access to a dictionary learned significantly more vocabulary words than those that were not provided with a dictionary. In addition, Hulstijn (1992) found that learners acquired more vocabulary when they inferred words themselves, as opposed to being given the definition. In a later study specifically designed to test Laufer and Hulstijn’s (2001) theory, Hulstijn and Laufer (2001) compared three groups with tasks including equal amounts of need, but with varying levels of search and evaluation.

Hulstijn and Laufer (2001) performed two experiments to empirically examine the effect of Task-Induced Involvement on vocabulary learning. In each experiment, 3 intact groups of advanced students of English in college participated. One experiment was conducted in Israel with 128 native Hebrew speakers, while the other was conducted in Holland with 97 native Dutch speakers. In both experiments, groups were randomly assigned to one of three task conditions. The tasks varied in terms of their Task-Induced Involvement. In the first task condition, participants read a text complete with a marginal gloss for 10 English vocabulary words and subsequently took a reading comprehension test. As need for the vocabulary words
was moderate in that it was necessary to complete the task, but search and evaluation were not present, this task had an involvement load of 1. In the second task, participants were required to both read the text and fill in 10 blanks with the appropriate vocabulary word from a word bank containing 15 English words and their translations, and then, subsequently, take a reading comprehension test. As this task additionally required the participants to evaluate the words, its Task-Induced Involvement was 2. In the third task, participants were to write a composition and appropriately include the target words, of which definitions and details about usage were provided, therein. This task was determined to have a task involvement load of 3, due to its strong evaluation in the need to determine multiple aspects of the words, syntactic and semantic, in order to use them appropriately. Either one week (Holland) or two weeks later (Israel), participants took a vocabulary test in which they were asked to provide the L1 translation or English definition of the ten target words. In the immediate and delayed vocabulary post-tests, Israeli participants assigned to task 3 were found to perform significantly better than those assigned to task 2, and those assigned to task 2 were found to perform significantly better than those assigned to task 1. For the experiment conducted in the Netherlands, however, while participants assigned to task 3 were found to perform significantly better in an analysis combining immediate and delayed post-tests, no difference was found between those assigned to task 2 and those assigned to task 1. When the scores of participants assigned to task 1 and task 2 were combined, the difference in performance between participants assigned to one of these two tasks and those assigned to task 3 was found to be significant for both the immediate and delayed vocabulary post-tests. Laufer and Hulstijn (2001) argue that their Task-Induced Involvement construct is strongly supported by the results of the experiment in Israel, as each increment in involvement load led to increased retention, but are only partially supported by the results of the
experiment in Holland, given the lack of a significant difference between those participants
assigned to tasks with an involvement load of 2 (task 2) and 1 (task 1).

In a conceptual replication of Hulstijn and Laufer (2001), Keating (2008) explores the
potential role of proficiency level, type of retention test, and time on task-induced involvement
loads. Seventy-nine students of third-semester Spanish were included in the final analysis.
Participants were assigned to one of three tasks. The first two tasks, as in Hulstijn and Laufer
(2001) required the participants to read a text and answer comprehension questions based on the
text, with the difference between these two tasks being whether the vocabulary information was
available in the glosses (task 1) or the students had to fill it in based on a list of words and
translations provided to them (task 2). Task 3 differed from Hulstijn and Laufer (2001) in that it
required participants to write a list of sentences with the vocabulary words as opposed to writing
a composition. Also, in addition to the L2 to L1 immediate and delayed retention test that was
also used in Hulstijn and Laufer (2001), Keating (2008) required participants to take an
immediate and delayed L1 to L2 retention test as well. In a two way mixed ANOVA with L2 to
L1 retention test score as the dependent variable, a main effect was found for time and task, with
the follow-up post hoc analysis revealing a significant difference between tasks 3 and 1 and
between tasks 2 and 1, but not between tasks 2 and 3. In the same type of analysis with L1 to L2
retention test score as the dependent variable, task and task and time interaction were found to be
significant, while time approached significance. In the post-hoc analysis of the one-way ANOVA
conducted on immediate post-test scores, participants assigned to task 3 were found to perform
significantly better than those assigned to task 2, and those assigned to task 2 were found to
perform significantly better than those assigned to task 1. In the post-hoc analysis of the one-way
ANOVA conducted on the delayed post-test score, however, the only significant difference was
found between those participants assigned to task 2 and those assigned to task 1. In their analysis of the time on task data, Keating (2008) found, when the scores were divided by the number of minutes spent on the task, that the differences between tasks on retention score were no longer significant.

Keating (2008) argue that, overall, the results of their study provide support for Laufer and Hulstijn’s (2001) Task-Induced Involvement construct. They note that their experiment extends the findings of Hulstijn and Laufer (2001) in that it found that higher task involvement load generally lead to higher retention at the low-proficiency level and for both active and passive retention tests. Their results also add nuance to Hulstijn and Laufer’s (2001) findings in that they found that gains in vocabulary knowledge did not hold as long for the active retention tests as they did for the passive retention tests. If tasks with a involvement load is indeed representative of processing levels, this particular result, if taken at face value, contradicts Craik and Lockhart’s (1972) postulation that traces developed from higher levels of processing are not only immediately stronger, but also more durable.

Nonetheless, a curious difference between Hulstijn and Laufer’s (2001) study and Keating (2008) may have played a role in the results, particularly the difference in gains between the active and passive retention tests. This difference is that participants were told, prior to beginning, that they were going to be tested on their vocabulary acquisition. While Keating (2008) emphatically argue that it is “highly unlikely” (p. 380) that this would have had an impact on the results, pointing out a comment by Hulstijn (2001) that “word learning is intentional only if learners make a conscious effort to commit new words to memory” (as cited in Keating, 2008, p. 380), it should be noted that Keating (2008) did not collect survey data to probe for this very possibility. Indeed, the only data that Keating (2008) provides to support his assurance that such
a possibility is highly unlikely is indirect evidence, in the form of time data and perceptions of the researcher. Even if it is accepted that for any individual participant it is unlikely that he intentionally learned the vocabulary words, certainly it is less so that at least some participants may have engaged in the behavior. Generally speaking, one might speculate that a potential impact of participants intentionally committing words to memory is that it may have reduced the differences between participants assigned to different task loads by adding an additional task load, in the form of need, to each, and, in turn, reducing their relative differences in task load. Given that Keating (2008) concludes that his experiment supports the Task-Induced Involvement construct, this is not a particularly concerning impact. However, a more specific potential impact is in regards to the difference that Keating (2008) found between the active and passive retention delay post-tests, with active post-tests showing less differences in gains across tasks. One might speculate that an intention to commit to memory on the part of some participants may have increased active test scores overall and diluted the potential impact of higher task loads.

Each of the previously explored depth of processing studies explored relies on tasks alone to assign levels of processing (Gass et al., 2003; Hulstijn & Laufer, 2001; Keating, 2008; Shook, 1994). While task differences may indeed make different levels of processing more likely, without collecting protocols or survey data from the participants, the researchers cannot be certain that the participants processed at the level to which they were assigned.

Qi and Lapkin (2001) examined the role of depth of processing, or what they refer to as the “quality” (p. 291) of noticing on writing improvement using concurrent verbal reports. In the first stage of the experiment, two ESL students, one with a high level of proficiency and one with a low level of proficiency, first wrote a story based on a drawing of a crime scene that they were given. In the second stage, the ESL students were given a reformulated (corrected) version
of their stories together with their original drafts and were asked to Think-Aloud while they
compared them. Unlike in Shook (1994), Qi and Lapkin (2001) operationalized depth of
processing by examining these Think-Alouds, and found two levels of processing of the
corrections made in the reformulated version: “noticing without giving reasons” (p. 291) which
can be seen as a low level of processing, and noticing with giving a reason, which can be seen as
a high level of processing. At stage three, the participants rewrote their original drafts without
the aid of the reformulations they reviewed in stage 2. Qi and Lapkin (2001) observed that level
of processing (“quality” of noticing) appeared to be related to writing improvement, with items
processed at the high level of processing at stage two (i.e., noticing with a reason) leading to
more corrections at stage three when compared to items processed at a low level of processing
(i.e., noticing without a reason). Needless to say, given that only two participants were involved,
caution is needed in the interpretation of these findings.

In a post-hoc analysis, Leow et al. (2008) explored the impact of depth of processing for
form on comprehension. As in Qi and Lapkin (2001), Leow et al. (2008) operationalized depth of
processing by examining the Think-Aloud data. Leow et al. (2008) established three processing
levels: (I) Only circling; (II) Circling and reporting; (III) Circling, reporting and interpreting or
translating. They found no consistent link between level of processing and comprehension score,
and concluded that the data did not appear to show that higher levels of processing and
comprehension scores are related. Leow et al. (2008) suggested, however, that caution is needed
in interpreting the results due to the low numbers in experimental cells.

Morgan-Short et al. performed a conceptual replication of Leow et al (2008). They re-
examined the potential relationship between level of processing and level of comprehension with
the inclusion of many more participants and explored the possibility that reactivity might have
played a role in the results. Three hundred and sixty-one final participants followed the same procedure as in Leow et al. (2008), except that 187 were assigned to a Non-Think-Aloud condition. Morgan-Short et al., as in Leow et al. (2008), eliminated those participants in the Think-Aloud condition who went back to the text as they were answering the multiple-choice questions, or what Leow et al. (2008) refer to as backtracking. However, unlike in Leow et al. (2008), Morgan-Short et al. (2012) also eliminated those participants that “reread the text for comprehension,” what they refer to as backtracking, because, in their view, these participants could not be seen as “having [simultaneously processed for both] form and meaning” (p. 672). To avoid confusion, in this thesis, the action of going back to the text to answer the multiple-choice questions will be referred to as backtracking, going back within the text to read it again will be referred to as rereading. In total, Morgan-Short et al. eliminated 53 participants, or 25% of the original size of the Think-Aloud group, for rereading. Morgan-Short et al., however, did not eliminate the individuals in the Non-Think-Aloud condition (not present in Leow et al., 2008) who reread. Morgan-Short et al. did find the Think-Aloud condition to have a significant impact on the results, although they emphasize that the effect size found was too minimal ($n^2 = 0.01$) to be of practical value. From the article, it is not possible to determine the extent to which participants reread in Morgan-Short et al. (2012), and, therefore, it is difficult to evaluate their claim that such participants did not simultaneously process for both form and meaning.

Regardless, however, making these eliminations in the Think-Aloud group, but not in the Non-Think-Aloud group, may have played a role in the results.

As in Leow et al. (2008), no effect was found for Condition. No interaction was found between the Think-Aloud condition and the Attentional condition, suggesting that the minimal effect Morgan-Short et al. found for thinking-aloud did not provide an advantage or disadvantage
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

to any of the Attentional conditions, although this too must be interpreted in the context of the aforementioned compromise in the comparison between the Think-Aloud and Non-Think-Aloud groups.

Morgan-Short et al. also explored the effect of depth of processing on the target form, coding the Think-Alouds for three levels of processing, as in Leow et al. (2008). They found that there was a significant positive correlation between the depth of processing level and comprehension score, and that the correlation had a medium effect size. According to Morgan-Short et al., this result “would not seem consistent with” (p. 680) VanPatten’s (2004) Primacy of Meaning Principle, which “might have [] predicted” that “attending to grammatical or lexical form while reading a L2 text for meaning” would be detrimental to performance. (p. 683).

In summary, studies to date on depth of processing have come to diverging results with regards to their impact on performance (Gass et al., 2003; Hulstijn & Laufer, 2001; Keating, 2008; Leow et al., 2008; Morgan-Short et al., 2012; Shook, 1994; Qi & Lapkin, 2001). Two studies to date have found no evidence of depth of processing impacting performance (Leow et al., 2008; Shook, 1994), while five found at least some evidence of depth of processing having an impact (Gass et al., 2003; Hulstijn & Laufer, 2001; Keating, 2008; Morgan-Short et al., 2012; Qi & Lapkin, 2001).

Methodological Issues of Empirical Research

One of two general methodological concerns, among other more specific issues, some of which have been previously pointed out, can be observed for each of these studies. While Gass et al. (2003), Hulstijn and Laufer (2001), Keating (2008), Shook (1994) are methodologically flawed in that they relied on tasks alone to determine processing levels, and therefore lack
evidence that participants processed at the level they were assigned to, Qi and Lapkin (2001), Morgan-Short et al. (2012), and Leow et al. (2008), in their post-hoc analyses, also have a collective, and largely opposite, methodological issue. These studies (Leow et al., 2008; Morgan-Short et al., 2012; Qi & Lapkin, 2001) coded for processing level entirely based on the Think-Aloud data, and, therefore, did not randomly assign participants to processing levels, potentially allowing for mediating variables to play a role. It should be noted, however, that Qi and Lapkin (2001), unlike Morgan-Short et al. (2012), coded each item, as opposed to each participant, for depth of processing, allowing for a partially within-subject comparison that may have reduced the impact of mediating variables. This methodological issue, among others more specific to Morgan-Short et al. (2012), and the study they replicated (Leow et al., 2008), will now be discussed.

One concern, given the possibility that reactivity may be more likely for more cognitively complex tasks, as pointed out by Morgan-Short et al. (2012), is that the Think-Aloud group was either not controlled for (Leow et al., 2008) or that the control group was compromised (Morgan-Short et al., 2012).

Second, in Morgan-Short et al., as well as in the study they replicated (Leow et al., 2008), unlike in previous studies (e.g., VanPatten, 1990; Greenslade et al., 1999; Wong, 2001), no time limit was placed on participants. Instead, Leow et al. simply instructed participants to “read as quickly as” (p. 677) they could, while Morgan-Short et al. (2012) did not provide participants with instructions regarding their pace. In the case of Leow et al. (2008), however, one might speculate that participants may not have followed these instructions given that no additional method was used to encourage participants to proceed speedily or verify that they did indeed proceed as quickly as possible. Given the lack of instructions regarding pace, it is potentially
even less likely that participants consistently completed the reading as fast as they could in Morgan-Short et al. (2012). This may have had an impact on the results. Indeed, Morgan-Short et al. (2012) points out that timing may have played a role in differences between the written and aural modality that have been found in previous research on processing for form and meaning simultaneously, as “attentional constraints are more apparent in the written mode[…] when the timing of input is controlled” (p. 679). In other words, as one crucial aspect of VanPatten’s (2004) input processing model is that learners will have difficulty processing for form and meaning simultaneously when cognitive resources are limited, in the event that participants in the processing for form conditions did indeed take additional time to compensate for the additional processing task, this could have given them precisely the cognitive resources they needed to process for form and meaning without having any detrimental effect on either, a criticism that was also made previously in this paper with regards to the Form and Meaning condition of Hulstijn (1989). Indeed, in studies in which a time-limit was used, unlike in Leow et al. (2008) and Morgan-Short et al. (2012), processing for form was found to negatively impact comprehension (Greenslade et al., 1999; VanPatten, 1990; Wong, 2001). Moreover, Leow et al. (2008) found time to be near significance in a one-way ANOVA, a seemingly inconclusive result. An additional instrument, therefore, may be warranted in order to ensure that participants are proceeding as quickly as possible.

Third, Morgan-Short et al.’s (2012) coding scheme for processing levels II and III, also used in the study they replicated (Leow et al., 2008), may have been methodologically flawed for two reasons. First, there may not have been a difference in terms of processing between participants coded as level I and participants coded as level II based on their coding criteria. Morgan-Short et al. (2012) coded participants for processing at level I if they only circled the
forms, and at level II if they exhibited one of three different criteria in the concurrent verbal reports: “(a) pronouncing the target form … (b) noticeably pausing immediately before or after the target form … or (c) making a non-metalinguistic comment about the form” (Morgan-Short et al., 2012, pp. 672-673). Given that the forms used in Morgan-Short et al. (2012) were grammatical, however, and not phonological, there is no apparent difference between a participant circling the forms and pronouncing them in terms of the strength of the connection between the form and its meaning. Second, Morgan-Short et al.’s (2012) coding only required that participants process one form item at level II or III to be coded as having processed at the respective level. While Morgan-Short et al. (2012) do not include data regarding the percentage of forms that participants processed at the levels to which they were coded, insight can be gained from the study that Morgan-Short et al. (2012) replicated (Leow et al., 2008). Using the same coding criteria as Morgan-Short et al. (2012), Leow et al. (2008) found, for the group coded as processing at level III, that “the majority of participants who translated or interpreted the target forms did so on very few items …” (p. 681). As participants were tested on their comprehension of the entire passage, and not just the paragraphs of the passage where they processed forms at a higher level, it would seem that a more consistent high level of processing than that used in Morgan-Short et al. (2012) would be necessary in order to accurately determine its impact on comprehension. In other words, it may be that amount of processing, in addition to depth of processing, might have an impact on comprehension.

One final methodological issue for Morgan-Short et al. (2012) and the original study (Leow et al., 2008) is that processing level was coded for after the studies were conducted as opposed to being randomly assigned to participants before they began the tasks. As a result of this methodology, it is possible that overall more capable participants may have been more likely
both to interpret forms and to comprehend the text better, which may explain why the experiment found a positive relationship between processing at a higher level and higher overall comprehension. That is to say, without random assignment to a variable X, other variables that may be both related to variable X and higher performance in an assessment task Y, known as mediating variables, can lead to erroneous results that suggest that X is related to Y.

Working memory capacity, for example, which has been found to be related to the assessment task in studies on simultaneously processing for form and meaning (Greenslade, Bowden, & Sanz, 1999; Leow, Hsieh, & Moreno, 2008; Morgan-Short, Heil, Botero-Moriarty, & Ebert, 2012; VanPatten, 1990; Wong, 2001), namely reading comprehension (e.g., Harrington & Sawyer, 1992; Walter, 2004), could potentially have confounded the results in studies on processing for meaning and form. To this end, controlling for level of processing before participants engage in the reading process clearly warrants further investigation from a methodological perspective.

Working Memory

Given the description of the nature of the limit of input processing, that being that when cognitive resources are exhausted input cannot be fully processed, it logically follows from this theory that individuals with higher levels of working memory capacity may be able to perform better on exhaustive input processing tasks.

In the following sections, working memory will be discussed both in theoretical and practical terms, as well as both in cognitive science and language acquisition.
Working Memory in Cognitive Science

In psychology, as outlined in Conway, Jarrold, Kane, Miyake, and Towse (2007), the limitations of short-term memory have been examined for over a century now. Ebbinghaus (1885/1964) investigated memory by reading several series of nonsense syllables aloud, each with a different number of total syllables, and, subsequently, attempting to recite them himself. Ebbinghaus (1885/1964) controlled for both the time he spent reading the series and avoided stressing his voice. He found that he was able to memorize 7 syllables after just one reading, while 12 syllables required 16.6 readings on average, and 16 syllables required 30.0. The exponential increase in the number of reading times necessary for additional syllables beyond seven is anecdotal evidence of an upper limit to short-term memory. Jacobs (1887) empirically examined short-term memory by using Ebbinghaus’ (1885/1964) method with children and young-adults. In his study, participants between 8 and 20 years old were asked to, among other tasks, attempt to memorize a series of nonsense syllables. He found that span increased with both age and performance in the classroom. Eleven-year-old schoolgirls, for example, memorized an average of 5.3 syllables, while their 20-year-old counterparts memorized 7.0.

Short-term memory span tasks, however, were limited in that they did not take into account “normal human processing” (Baddeley & Hitch, 1974, p. 47), that is, the way memory is used with all the distractions and tasks of everyday life. Baddeley and Hitch (1974) note that while short-term memory had been “assumed to be responsible” for cognitive tasks such as “problem solving, language comprehension and long-term learning,” little empirical evidence had been gathered to demonstrate such claims (pp. 47-48). Baddeley and Hitch (1974) believed that this type of practical memory could be taken into account by making participants complete complex cognitive tasks and store old information at the same time. In their first experiment,
Baddeley and Hitch (1974) showed 24 participants either one or two digits, which they subsequently were asked to retain while they performed a reasoning task. Baddeley and Hitch (1974) found no difference between the two groups with regards to the time participants took to complete the reasoning task, leading them to speculate that perhaps a heavier load would be necessary to see a detriment in performance. This speculation was confirmed in their second experiment, which increased the load to six digits and asked 24 participants either to focus on recalling the digits or to attempt to both recall the digits and complete the reasoning task as accurately as possible. Baddeley and Hitch (1974) found that those participants who focused on recalling the digits did recall more digits, but also took longer on the reasoning task, leading Baddeley and Hitch (1974) to theorize a trade-off between what they called reasoning and recall (later known as processing and storage, respectively). Working memory, as it is known today, was essentially captured by Baddeley and Hitch’s (1974) original account, later defined as the ability to simultaneously “maintain and process goal relevant information” (Conway et al., 2007, p. 3). Empirical support for Baddeley and Hitch’s (1974) theory of a distinction between memory itself and memory while processing was reported in Engle et al. (1999). In their study, 133 participants’ scores on three traditional short-term and working memory tasks showed that while working memory capacity and short-term memory were highly correlated, the six tasks were best fit by a two-factor model that separated the two types of memory tasks. That is to say, while sharing the memory component, short-term memory and working memory capacity are still two different constructs.
Three of the earliest working memory span tasks developed were Daneman and Carpenter’s (1980) Reading Span and Listening Span Tasks, and Turner and Engle’s (1989) Operation Span Task.

In Daneman and Carpenter’s (1980) original study, the researchers employed three different Reading Span Tasks (RSPANs) and one Listening Span Task (LSPAN). In the RSPANs, the basic task was to read several series of sentences, and, at the end of each series, recall the last words of the sentences presented to them in that series. The RSPANs differed in that in two of the three RSPANs participants were given the additional task of stating whether each sentence was true or false. Additionally, one of these three RSPANs was silent, meaning the participants were to read the sentences silently. The LSPAN that the authors employed was similar to the RSPANs with the true/false component, except that the sentences were read to the participants, as opposed to the participants reading the sentences themselves. Requiring the participants to state whether each sentence was true or false, as the authors state, has the advantage of ensuring that the participants are processing the entire sentence and not just the last word. All RSPANs and the one LSPAN were found to be significantly related to reading comprehension in Daneman and Carpenter (1980).

Turner and Engle (1989), arguing that working memory capacity should be related to higher cognitive functioning regardless of whether or not the cognitive task and the concurrent processing task of the working memory capacity span are related, developed the Operation Span Task (OSPAN), which replaced the sentences that were used in Daneman and Carpenter’s (1980) Reading Span Task with mathematical operations. In Turner and Engle (1989), participants completed two OSPANs. In one OSPAN, participants were required to determine the correctness
of a series of mathematical operations, and, after each series, recall the numerical digits that appeared after each operation, while in the remaining OSPAN participants were required to recall words instead of numerical digits. Turner and Engle (1989) found the reading span task and the operation span task to be strongly correlated. Together, RSPANs and OSPANs have been found to be significantly related to a number of cognitive tasks, many of them language related, such as reading comprehension (e.g., Walter, 2004), programming skill acquisition (Shute, 1991), reasoning (Barrouillet, 1996), as well as vocabulary and grammar learning (Martin & Ellis, 2012).

Kane et al. (2004) and Shah and Miyake (1996) have developed spatial working memory span tasks. Shah and Miyake (1996) theorized that spatial and verbal working memory spans may be assessing different working memory resource pools. In Shah and Miyake’s (1996) Rotation Span Task, participants were presented with sets of English capital letters and their mirror images. These capital letters were also facing a particular direction. Participants were asked to determine whether the letter was normal or a mirror image of a normal English letter as the letters were presented to them. They were then asked to recall the orientation of the letters after each set. Shah and Miyake (1996) found the Rotation Span Task, but not the Reading Span Task, to be significantly related to participants Composite Spatial Score. They also found the Reading Span Task, but not the Spatial Span Task, to be related to Verbal SAT score, providing evidence in favor of their hypothesis that spatial working memory is distinct from verbal working memory. These results suggest that spatial working memory tasks, while perhaps useful in their relationship to spatial tasks, as Shah and Miyake (1996) note, should be avoided in SLA research. In Kane et al.’s (2004) Symmetry Span, participants were shown sets of 8x8 matrices with some squares filled in black and were asked to determine whether or not the squares
constituted a symmetrical design. After each 8x8 matrix, participants were shown a 4x4 matrix with one square filled in red. Following each set, participants were required to recall the location of the red squares. Unlike in Shah and Miyake (1996), Kane et al. (2004) found spatial and verbal working memory capacity tasks to be significantly related and concluded that working memory capacity is domain-general, as opposed to domain-specific, suggesting that working memory capacity tasks, regardless of the domain of their components, may be interchangeable. Nonetheless, given that spatial working memory capacity tasks were not found to be related to reading comprehension in Shah and Miyake (1996), and that their relationship to language-related cognitive tasks has not been as extensively explored as the RSPAN and the OSPAN, for the moment, their use should be avoided in SLA studies that are not directly addressing their viability. In addition, the OSPAN is preferable to the RSPAN in studies on second language acquisition given that the processing component is not language-related.

The commonality of all working memory span tasks, such as those mentioned here, is that they each include a processing and storage component.

**Models of Working Memory**

While the basic components of processing and storage are generally agreed upon as being essential to working memory, the exact nature of this construct is still controversial, as Goo (2010) pointed out. Four basic models of working memory capacity have been developed to explain this important cognitive construct: the resource sharing account, the task-switching account, the executive attention view, and the inhibitory based executive control model.

According to Daneman and Carpenter (1980), working memory capacity can be thought of as “a trade-off between processing and storing functions” (p. 450). In other words, an individual can divert resources from either of these two functions of working memory
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

(processing new information or maintaining old information) in order to allocate more resources to the other. Daneman and Carpenter (1980) argue that information can be lost either due to “decay,” loss over time, or “displacement,” loss due to the incorporation of “additional structures” (pp. 450-451). If a task requires a lot of processing, less information will be able to be maintained. The resource-sharing account is supported by Baddeley and Hitch’s (1974) finding (described in more detail above) that participants who focused on storage processed more slowly. Daneman and Carpenter (1980) use reading comprehension as an example of where high working memory capacity individuals show an advantage. Engle’s et al. (1999) finding of a strong correlation between working memory capacity and short-term memory provides additional support for the resource-sharing model, in that it suggests there is a key storage component in working memory as well. Empirical support for the resource-sharing account was also reported in at least one result of Towse, Hitch, and Hutton (2002), which showed that recall was poorer in children when errors were made in the processing component of an addition OSPAN task. This indicates that higher levels of processing, which presumably occur when more errors are made, may be negatively influencing the children’s storage.

The task-switching account, first outlined by Baddeley and Hitch (1974), posits that working memory capacity is about more than just the resources one has available (Towse & Hitch, 2007). According to this model, working memory capacity is affected by the passage of time. That is, as time progresses, individuals forget old information when they begin to process new information. Previous studies that have found time spent processing to have a negative impact on recall in children provide support for this model (Towse, Hitch, & Hutton, 1998; Hitch, Towse and Hutton, 2001). Hitch et al. (2001), for example, found that children performed more poorly on both an operation span task and a reading span task when presented with longer
processing tasks at the end of the span tasks as opposed to at the beginning. Hitch et al. (2001) assert that the resource sharing model cannot account for this difference, as the total amount of processing throughout the span tasks was the same. While time may have been a factor, however, placing the longer processing tasks at the end of span tasks, at which point participants needed to be storing more information as well, arguably also placed a higher maximum cognitive burden on participants at any one point in the task, when compared to placing the longer processing tasks at the beginning, when participants were not required to store as much. This could potentially explain the lower results of the span tasks that placed the longer processing tasks at the end in terms of the resource-sharing account, as participants may have traded some crucial storage resources at the end of the spans to perform the longer processing tasks. In Towse, Hitch and Hutton (2002), the duration of the retention interval in an interpolated task, but not the difficulty of the processing component (multiplication vs. addition), was found to have a significant impact on recall in children, providing further evidence that time could be a crucial factor to working memory capacity. Towse and Hitch (2007) admit, however, that their account is limited, given that while the studies finding an impact of duration on performance were conducted on children, similar findings were not found with adult participants (Towse, Hitch, & Hutton, 2000).

The third model of working memory capacity, the executive attention view, posits that individual working memory capacity differences can be explained by differences in the ability to control attention (Engle, 2002, p. 20). A number of studies (e.g., Kane, Bleckley, Conway, & Engle, 2001; Kane & Engle, 2000) support this account of working memory. In Kane and Engle’s (2000) first experiment, for example, 216 participants were divided into high and low working memory capacity groups by an operation-word span task. All participants were then
randomly assigned one of two finger-tapping conditions that varied based on whether or not they provided an additional burden to participants. While tapping their fingers according to their condition, participants participated in three consecutive trials where they first saw a list of 10 words, and then were asked to recall the words. While the recall percentage for the first trial did not significantly differ between the two groups, during subsequent trials, the low working memory group was significantly less able to recall the new words than individuals with higher working memory spans under the no-load finger tapping condition. Under the load finger tapping condition, however, the two working memory capacity groups performed similarly. This showed that, without the additional burden of the load finger tapping condition, higher working memory participants were able to block out the proactive information of the first list, providing evidence in favor of the controlled attention view of working memory. Nonetheless, this study has aspects in common with a reading working memory span task, which has been shown to correlate with the operational working memory span task that Kane and Engle (2000) employed (Conway et al., 2005). Possibly, then, the relationship between working memory capacity and the task used in the experiment is due to this commonality, and not the ability to deal with proactive interference as the authors assert.

In another study, Kane, Bleckley, Conway, and Engle (2001), in two separate yet related experiments, selected the antisaccade task as the attentional control task, which greatly reduced any potential influence of “cognitive skill,” “surface similarity to a span task,” and “memory demands” (p. 181). In both experiments, participants, 203 in the first and 40 in the second, performed an antisaccade task where they fixated their vision on the middle of a display, and were asked to respond to target information that randomly appeared either on the right or left side of the display. Additionally, there were two conditions, a prosaccade condition that
presented an attention-capturing cue on the same side as the target information and an antisaccade condition that presented an attention-capturing cue on the opposite side. In both experiments, the high-working memory capacity group, as expected, had significantly lower latency times and higher accuracy when compared to the low-working memory group on the antisaccade condition. In the second experiment, which used eye-tracking, it was also shown that low working memory capacity participants more frequently moved their eyes to the cue and that they were slower in moving their eyes back to the target. The authors conclude that this study provides strong evidence that working memory capacity is related to attention control. In an interesting and unexpected result, in both experiments, while the two working memory group capacity groups did not perform differently on the prosaccade condition when it was given before the antisaccade condition, the higher working memory capacity group did perform significantly better when it was given after. This could indicate, as the authors suggest, that higher working memory capacity participants are flexible and better at adapting to new conditions.

The results of Kane and Engle (2000) and Kane et al (2001), however, while providing evidence that executive attention is one component of working memory capacity, fall short of providing evidence that executive attention is the central component of working memory, as Engle (2002) espouses. In each of the two studies, the authors compare high working memory capacity learners’ performance or latency in executive attention tasks to that of low working memory capacity participants, and find significant differences between the two groups. Only a minimal relationship between working memory capacity and executive attention, however, would be necessary to lead to these findings. A more convincing statistical analysis would have examined the strength of the correlation between working memory capacity and performance or
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

latency, which was addressed by Oberauer, Süb, Wilhelm, and Wittman (2003). In their study, Oberauer et al. (2003) developed a facet taxonomy of working memory tasks that included two content factors (visuo-spatial material and language and numerical material) and three functional factors (simultaneous storage and processing, supervision, and coordination). Supervision, also referred to as executive processes by Oberauer et al (2003), was defined as the “monitoring of ongoing cognitive processes and actions, the selective activation of relevant representations and procedures, and the suppression of irrelevant, distracting ones,” and coordination was defined as the ability to “construct new relations between elements” in order to “establish[] a mental structure on which the required response is based” (p. 169). One hundred and thirty-five participants performed tasks that combined one content factor and one functional factor, as well as six traditional working memory capacity tasks. Results showed that several content factor loadings were not significant, leading the authors to omit them in their analysis of functional factors. All of the functional factors, with the exception of supervision, were found to have a strong correlation to participant scores on the traditional working memory task scores. In the case of supervision, only half of the working memory span tasks were found to be significantly related to one of the two executive function indicators, and the highest of these correlations was 0.22, challenging the executive attention model’s central argument that working memory is essentially executive attention control.

The fourth model of working memory, the inhibitory-based executive control model, explains working memory as equivalent to three functions regarding information control: access, deletion, and restraint (Hasher, Lustig, & Zacks, 2007, p. 231). Access is the process of determining what information will gain access to an individual’s attention. To be efficient, the access function suppresses information that is irrelevant to goals. Dywan and Murphy (1996)
explored the role of access in reading comprehension. In their study, 28 older and 24 younger adults read a text either with or without distracters. Older adults were found to be more negatively affected by the distracters both in terms of time and reading stumbles. Additionally, while overall comprehension scores were not found to be significantly different between older and younger participants, older participants chose the distracter answer more frequently than their younger counterparts. Restraint is the ability to avoid strong responses, and is effective when it avoids strong responses that are not relevant to the task at hand. Older adults were found to have less restraint than their younger counterparts, showing significantly more priming effects for words they were told not to remember (May & Hasher, 1998). Significantly more restraint was also shown by participants with higher working memory capacity in an antisaccade task (explained above, Kane et al., 2001). Deletion is the ability to delete information not or no longer relevant to goals. Evidence showing the importance of deletion is found in Ham and Hasher (1992), where older adults, who have been shown to have lower working memory capacity (Salthouse & Babcock & Shaw, 1991), were less able to abandon their original inferences about a passage when they were presented with new information that disconfirmed them. Hasher compares the inhibitory-based executive control model with a spam blocker, where information that is not useful does not enter the system. This is the crucial distinction from Kane, Conway, Hambrick, and Engle’s (2007) executive attention model. Whereas Kane et al. (2007) view “excitatory mechanisms that activate representations” (Hasher et al., 2007, p. 241) as crucial to differences in working memory spans, Hasher et al. (2007) see working memory variation as primarily a function of inhibition.

Together, these models and their related studies provide a window into the nature of working memory. Given that studies have found results to support each one of them (cf. Engle et
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

al., 1999; Kane & Engle, 2000; May & Hasher, 1998; Towse et al., 2001), working memory
capacity can potentially be seen as represented partially by each of them. That is to say, working
memory capacity may be a factor of the processing and storage tradeoff, the passage of time, the
ability to control attention (similar to the access and restraint functions of Hasher et al., 2007),
and the ability to delete information that is no longer necessary. The complexity of working
memory capacity, combined with the relatively simple administration of working memory span
tasks, makes it an especially useful construct in research, especially in comparisons to other
abilities (Colom et al., 2010; Engle, 2001).

Indeed, the relationship between the ability to process and store information and higher
order cognitive tasks has been explored extensively in cognitive psychology (Engle, 2001). It has
been found that working memory capacity is related to bridge playing, “mental arithmetic,
complex reasoning, ability to follow directions, problem solving, bridge playing and spatial

Working Memory and Language

Working memory has been found to be related to a number of language-related cognitive
functions, such as programming skill acquisition (Shute, 1991), L2 vocabulary comprehension
(Kormos & Sáfár, 2008; Sunderman & Kroll, 2009), as well as L2 vocabulary and grammar
production and intake (Lesser, 2007; Martin & Ellis, 2012), L2 listening, speaking and reading
(Kormos & Sáfár, 2008), and L1 and L2 reading comprehension (Daneman & Carpenter, 1980;
In Shute (1991), the relationship between programming skill acquisition and working memory capacity, among other individual differences, was examined. In the study, 260 high school graduates attending San Antonio colleges and technical schools took a series of quantitative, verbal, and spatial working memory capacity tests, and a battery of tests on arithmetic reasoning, information processing speed, and general knowledge. Participants also completed 25 programming problems with the Pascal Programming Intelligent Tutoring System. Working memory capacity, the best predictor of computer programming skill acquisition, was the only factor to be independently significant when all other skills (arithmetic reasoning, information processing speed, etc.) were entered into the equation. Learning a computer language shares a number of commonalities with natural language learning, as it also involves learning grammar and vocabulary. More recent experiments have explored the relationship between working memory capacity and overall performance in the L2 (Kormos & Sáfár, 2008) as well as working memory capacity and vocabulary acquisition (Martin & Ellis, 2012; Sunderman & Kroll, 2009).

In Kormos and Sáfár (2008), the relationship between working memory and overall performance in the L2, including reading, speaking, writing, and listening, is examined. One hundred and twenty-one Hungarian high schoolers learning English participated in the experiment. Participants took a backwards digit span working memory test during their second year in high school. These scores were compared to scores that students received on the Cambridge First Certificate language exam, which they took for their final exam. Participants’ scores on three of the four skills, with the exception of writing, were found to significantly
correlate with participants’ scores on the backwards digit span, suggesting a relationship between working memory and L2 reading, speaking and listening.

Sunderman and Kroll (2009) explored the relationship between studying abroad, working memory capacity, and the ability to comprehend and produce L2 vocabulary items. Forty-eight participants, 34 of which had studied abroad, first completed a Reading Span Task. Subsequently they took a vocabulary Comprehension Measure, in which participants were asked to determine whether a series of word pairs, each with one Spanish and one English word, had the same meaning. Lastly, participants were required to complete a vocabulary-based Production Measure, in which participants were shown a series of pictures and asked to produce the Spanish name for each picture. Sunderman and Kroll (2009) found that, with study-abroad experience (SAE) factored out, working memory was significantly related to comprehension reaction time and accuracy, as well as marginally related to production accuracy. Sunderman and Kroll (2009) also found that for participants who recalled 24 words or less in the RSPAN (out of 80 total words), study-abroad experience had no effect on production accuracy, suggesting that a “threshold” working memory capacity is needed to benefit from the study experience (p. 10).

Both Sunderman and Kroll (2009) and Kormos and Sáfár’s (2008) studies, however, are limited as the researchers did not control for other potentially mediating variables such as learning outside of the classroom. For example, it is possible that participants with higher working memory capacities tend to study more, which would have confounded the results. In order to address this issue, Martin and Ellis (2012) looked at the relationship between working memory capacity and grammar and vocabulary acquisition in a laboratory setting with an artificial language.
Martin and Ellis (2012) empirically investigated the relationships between vocabulary learning, grammar acquisition, and working memory capacity. In the first session of the experiment, 40 participants from a large American University completed an LSPAN, in addition to two Phonological Short-Term Memory (PSTM) tests. Subsequently, they listened to and practiced individual words and sentences in an artificial language, and were provided with a written English translation and a relevant illustration. Participants then practiced translating into the artificial language with the English translation and illustration. In the second session, participants first practiced the artificial language similar to in the first session, and then took a test for which they were required to translate into (production score) and from (comprehension score) the artificial language. In this test, 26 sentences were repeated from the practice sessions and 24 were completely novel. Martin and Ellis (2012) found that vocabulary production, but not vocabulary comprehension, was positively correlated to working memory capacity. Martin and Ellis (2012) also found grammatical accuracy on novel sentences to be related to working memory capacity. It is important to note, however, that when vocabulary score was factored in, while overall grammatical accuracy was found to be significant, grammatical accuracy on novel sentences was no longer significant, suggesting that grammatical advantages for higher working memory capacity language learners may be related to their ability to memorize vocabulary and repeated sentences.

As in Martin and Ellis (2012), Leeser (2007) also empirically examined the relationship between grammar acquisition and working memory capacity. In the experiment, 146 high-beginner L2 Spanish learners were divided into two groups. One group read two familiar passages, and the other group read two passages on unfamiliar topics. All participants also took a form recognition test in which they were to identify the 19 target future verbs which they saw in
the text, and a tense identification task in which they were required to translate 20 Spanish noun and corresponding verbs. While the form recognition test was scored in a way that corrected for guessing and “considers the proportion of correct and incorrect answers,” the tense identification task only considered the translations of verbs in the future tense (p. 244). Results showed a main effect for both working memory capacity and topic familiarity, as well as the interaction between working memory capacity and topic familiarity. On the tense identification task, however, while a significant effect was found for topic familiarity, no such effect was found for working memory capacity or interaction between working memory capacity and topic familiarity. The lack of a significant finding for working memory capacity on the tense identification task should be interpreted with caution for two reasons. First, the scoring method used limited the total number of items to 5. Second, it is possible that some astute participants, who presumably were exposed to present, past and conditional, given the authors stated reason for using the future tense (“to employ a form to which the learners at the beginning level had not previously been exposed”, p. 240), were able to deduce that the one tense that they did not recognize was the remaining future tense.

Working memory capacity has also been found to be related to general fluid intelligence and multitask performance (Buhner et al., 2006; Colom et al., 2010; König et al., 2005; Unsworth, 2010). The relationship between working memory and both L1 and L2 reading comprehension, as well as the relationship between WM and multitask performance are most relevant to the ability of participants to process for form and meaning simultaneously.
Working Memory and L1 Reading Comprehension

Daneman and Carpenter (1980) in two separate experiments, Engle, Cantor, and Carullo (1992), and Walter (2004) found L1 reading comprehension to be related to working memory capacity. In Daneman and Carpenter’s (1980) first experiment, 20 college undergraduates first took a reading span task to determine their working memory capacity. The reading span task required participants to read several series of unrelated sentences with a different number of sentences in each series, and, at the end of each series, recall the last words of the sentences presented to them in that series. After completing the reading span task, participants completed a series of reading comprehension tests. In the reading comprehension tests, participants read 12 passages, and, after each passage, were asked two reading comprehension questions, one of which inquired about the referent of the last sentence of each passage, and the second of which asked about another piece of information in the passage. Lastly, participants took a short-term memory word span test. Working memory capacity, as determined by the reading span task, was significantly related to both types of reading comprehension questions, while short-term memory was not found to be significantly related to either. In Daneman and Carpenter’s (1980) second experiment, 21 participants took an oral reading span, a silent reading span test, and a listening span test. The oral reading span test, the silent reading span test, and the listening span test required participants to read aloud, read silently, or listen to, respectively, a series of sentences and determine whether each is true or false. As in experiment one, participants recalled the last words of the sentences after each series. Subsequently, participants took the reading comprehension tests, which were identical to those in experiment 1 except that participants answered 4 questions, of which 2 were fact-based, 1 was referent-based, and 1 asked participants to name the title of the passage. All three of the working memory span tasks were found to be
significantly related to participants’ scores on fact-based questions and participants’ scores on referent-based questions. Nonetheless, all of Daneman and Carpenter’s (1980) WMC span tasks were verbal in that they required participants to read or listen to sentences. Therefore, it may be the case that the common verbal element of reading sentences in both the WMC span tasks and the reading comprehension tasks is responsible for the correlation that Daneman and Carpenter (1980) found between working memory capacity and reading comprehension.

Engle, Cantor, and Carullo (1992) examined the relationship between participants’ performance on two self-paced working memory capacity span tasks and their verbal SAT (VSAT) scores in two separate, yet related, experiments. In the first experiment, 70 participants first took a self-paced Operation-word span task (OSPAN). As in traditional OSPANs, participants solved several series of operations, with each operation being followed by a word. After each series of operations, participants then needed to recall these words. In contrast to traditional OSPANs, however, participants were shown only one part of 5-part mathematical operations at a time in order to allow the researchers to later determine how much time each portion of the OSPAN was viewed. After taking the self-paced OSPAN, participants took an OSPAN without recall words. No significant differences were found for accuracy or latency on the OSPAN without storage between high and low WMC participants. On the contrary, for the OSPAN with recall words, VSAT score was found to significantly correlate with higher working memory capacity, demonstrating the importance of the processing element of working memory capacity. And, although high WMC participants were found to spend significantly more time on the recall words than low WMC participants, VSAT scores were still found to significantly correlate with high working memory capacity even when time spent viewing the recall words was partialed out, leading the authors to conclude that higher working memory capacity
participants are more than just better strategizers than their lower working memory capacity participants. The finding of a relationship between a non-verbal working memory capacity task (OSPAN) and reading comprehension provides crucial support to Daneman and Carpenter’s (1980) finding of a relationship between WMC and reading comprehension, as it cannot be explained away as being a byproduct of similarities between the two tasks. In the second experiment, similar to the first experiment except that a Reading span task was used, which, as in the case of the OSPAN, showed participants pieces of a sentence at a time, working memory capacity was again found to be related to the VSAT, even when time spent viewing the recall word was partialed out. In both experiments, therefore, Engle et al. (1992) demonstrate that the relationship between working memory capacity and reading comprehension, while potentially in part due to higher working memory capacity participants employing better strategies, such as spending more time on the recall, is, nonetheless, also indicative of the increased ability of higher working memory capacity learners to actively maintain information while simultaneously engaging in an additional task. Subsequent studies have also found the relationship between working memory capacity and reading comprehension to be applicable to speakers, both native and non-native, of languages other than English (Harrington & Sawyer, 1992; Leeser, 2007; Walter, 2004).

Working Memory and L2 Reading Comprehension

Walter (2004), Harrington and Sawyer (1992), and Leeser (2007) found L2 reading comprehension to be related to working memory capacity, each with different measures and assessment tasks for reading comprehension. Chun and Payne (2004), however, did not find a relationship between working memory and reading comprehension.
In Walter (2004), 41 French speakers learning English read 8 passages in English and 8 passages in French and completed a summary completion task after each text. In addition, participants took a verbal WM span task in both French and English. The verbal WM span tasks included a processing component of identifying whether sentences were logical or not, and a storage component requiring them to recall the last words of these sentences. Walter (2004) calculated one overall WM score by averaging English and French processing score, recall score, and reaction time. The study found that overall WM was significantly correlated to both French and English comprehension score, providing evidence that working memory capacity is related to reading comprehension, even when reading in a foreign language. In Harrington and Sawyer (1992), 34 native Japanese speakers learning English were first administered a digit span task, a word span task, and a reading span task. After these memory tasks, participants were given a cloze test, and their previous TOEFL reading and grammar scores were examined. Harrington and Sawyer (1992) found working memory capacity, as measured by the reading span task, was strongly correlated to both the TOEFL reading and grammar scores (.57). The correlation between the short-term memory tasks, as measured by the digit and word tasks, and the reading and grammar scores, while still significant, was also significantly less strong (.23), indicating that WMC was a better predictor of reading comprehension ability.

In Leeser (2007), the relationships between reading comprehension and both topic familiarity and working memory capacity were explored. As detailed earlier (Working Memory and Second Language Acquisition, p. 49), 146 high-beginner L2 Spanish learners were divided into two groups, with one group reading two familiar passages, and the other group reading two passages on unfamiliar topics. Following each of these passages, participants completed a comprehension recall protocol in their L1. Participants also took a Reading Span Task. Leeser
(2007) performed a 2x2 ANOVA and found topic familiarity and working memory to be significant factors, with interaction between topic familiarity and working memory approaching significance. This supports previous studies cited here that found working memory to be a significant predictor of reading comprehension score (e.g., Walter, 2004). In addition, given the inclusion of topic familiarity, the results suggest that working memory capacity may be an important individual variable for reading comprehension with both familiar and unfamiliar texts. Nonetheless, the results in this particular study may need to be interpreted with caution, given that the assessment task, the comprehension recall protocol, in which the author himself notes that participants were to “write down all they could remember” (p. 244), could have potentially favored higher working memory capacity participants. Chun and Payne (2004) is an exception to this and other studies explored here finding a relationship between reading comprehension and working memory.

Chun and Payne (2004) explored the relationships between reading comprehension, working memory, and look-up behavior. Thirteen college students taking a second-year German course participated in the study. On the first day, participants read a short story using a multimedia program that allowed them to look up as many words as they wanted. They also watched a video preview of the story, and were permitted to listen to an audio narration of the story. While still having access to the story, participants completed a series of multiple-choice questions, and drag-and-drop comprehension questions. For the drag-and-drop comprehension questions, participants were required to select the portion of the text that answered the question. On the second day, participants were allowed to reread the study and then wrote a recall protocol. Participants also were required to take a Reading Span Task. Chun and Payne (2004) did not find a relationship between reading comprehension and working memory capacity.
A number of methodological issues, however, should be considered when interpreting this result. As participants were allowed to access the text at the same time that they completed drag-and-drop questions, this question-type appears to be assessing the ability of participants to find answers in a text more than comprehension per se. Second, given that participants also watched a video preview of the story, were given the choice of listening to an audio narration or not, and were allowed to look-up however many words they wanted, it is plausible that any number of these factors could have had an impact on the result. To speculate, lower working memory capacity participants could have opted to listen to the audio narration more frequently than higher working memory capacity participants because they felt that they needed the reinforcement more. Lastly, the experiment was conducted with only thirteen participants, which, given the inclusion of two variables, may have been too few to obtain significant results. Chun and Payne (2004) also noted in their conclusion that future research may want to include more participants.

Given that reading comprehension and working memory capacity have been found to be related in a variety of studies which employed different working memory capacity span tasks (Engle et al., 1992), different languages (Harrington & Sawyer, 1992; Walter, 2004), both familiar and unfamiliar texts (Leeser, 2007), and also explored the potential effect of viewing time (Engle et al, 1992), and, affirmed the strength of the relationship between working memory capacity and reading comprehension, the relationship between reading comprehension and working memory capacity appears to be well-established. Perhaps even more relevant, however, is the finding that working memory is strongly related to multitask performance (Bühner, König, Pick and Krumm, 2006; Colom et al., 2010; König et al., 2005), given that the current proposal involves the “dual task” of processing for meaning and form simultaneously (Morgan-Short et
al., 2012, p. 666), and, for some participants, the multitask activity of doing so while thinking aloud at the same time.

Working Memory and Multitask Performance

König et al. (2005) was the first study to explore the relationship between working memory capacity and multitask performance. König et al. (2005) theorized that working memory could be a potential predictor of multitask performance as the construct involves the simultaneous use of storage and attention, which are both arguably necessary when individuals are multitasking (p. 245). In addition, the potential roles of fluid intelligence, attention, polychronicity (preferring multitasking activities), and extraversion were examined. In the study, 131 students of a German University first took a socio-demographic questionnaire, and the Inventory of Polychronic Values survey (p. 252). Second, participants completed the Simultaneous capacity/Multitasking scenario known as the SIMKAP. In the SIMKAP, participants needed to divide their attention between two tasks, the first of which required participants to pay attention and respond to a specific detail, marking identical numbers across two columns, and the second of which required participants to read and use a filled-out calendar by answering questions based on it. After performing the SIMKAP, participants took the Test Battery for Attentional Performance, which was then followed by three WM span tasks, namely, the Reading Span test, the Spatial Coordination test, and the Switching Numerical test. Participants completed the experiment by taking the computerized version of the Intelligenz-Struktur-Test 2000R, a fluid intelligence test. König et al. (2005) found that only working memory, attention, and fluid intelligence predicted multitask performance on the SIMKAP, and that working memory was the strongest of these predictors. Indeed, fluid intelligence only
explained 3 to 5% of additional variance after working memory in the hierarchical multiple regression analysis. König et al. (2005) suggested that future research explore the specific aspects of working memory capacity that contribute to its prediction of multitask performance.

Bühner et al. (2006) followed up on König et al.’s (2005) suggestion by comparing participants’ performance on a multidimensional model of working memory (as developed by Oberauer et al., 2003, and mentioned above) with their scores on the SIMKAP. One hundred and thirty-five students at a German University participated in the study. In the first session, participants first took the Go/No-Go attention test and the Divided Attention Test (for a description of both, see Zimmermann & Fimm, 2002). Subsequently, participants worked on three types of working memory capacity tasks: Storage in the context of processing tasks; Coordination tasks; and Supervision tasks. The Storage in the context of processing tasks, similar to the working memory capacity tasks used in König et al. (2005), required participants to store different types of information (verbal, numerical, dots, figures), while they performed another task, such as determining whether or not a pattern of dots is symmetrical. The Coordination tasks required participants to monitor changing items while simultaneously determining relationships between them. The verbal coordination task, for example, showed participants a 3x3 matrix with a word in each cell, and required them to detect when three rhyming words formed a line. In the Supervision tasks, participants applied decision rules, such as plant vs. animal and monosyllabic vs. bisyllabic, to stimuli. These decision rules sometimes remained the same (no-switching) and sometimes changed (switching). In the second session, participants took the SIMKAP (as described in the discussion of König et al., 2005. above). In the multiple regression analysis of the data, Bühner et al. (2006) found that both attention and working memory were predictors of multitasking performance on the SIMKAP, and that working
memory accounted for the majority of the variance in multitasking score. Of the working memory tasks, only Storage significantly predicted SIMKAP error score, that is, participants’ performance on the primary task (marking the matching number), while Coordination statistically predicted SIMKAP question score, that is, performance on the secondary task (answering questions based on a given calendar with appointments). Supervision did not significantly predict either of the two scores. Given the differential predictive values of Storage and Coordination, Bühner et al. (2006) theorize that “errors in multitasking might arise due to a low ability to store information while processing another task” (p. 271). Also due to these disparate predictive values, Bühner et al. (2006) argue that the tridimensional model of working memory capacity is superior to traditional WM span tasks. Nonetheless, Bühner et al.’s (2006) scoring of the Storage task likely had an impact of the diminished predictive power of this task with regards to the secondary task of the SIMKAP. As Colom et al. (2010) pointed out, Bühner et al. (2006) only scored the storage portion of the Storage task. Additionally, Bühner et al. (2006) did not establish any baseline for eliminating participants that did not perform well on the processing portion of the task. Not including processing as a basis for scoring or eliminating participants based on processing score may have resulted in the inclusion of some participants who were not processing or who were processing very little, potentially explaining why the Storage Task only explained SIMKAP error score, and not SIMKAP question score, which requires processing.

Colom et al. (2010) empirically examined the predictive power of general fluid intelligence, as well as both the processing and storage components of working memory capacity tasks, on multitask performance. In the study, “317 applicants for air traffic control training courses” first took the TRASI intelligence test. Participants then worked on the divided attention
and funnel tasks measuring multitasking performance (p. 544). In the Divided Attention Task, participants were asked to guide two dots on a screen towards their target while simultaneously pressing a space bar when the dots turned a specific color. Score on the Divided Attention Task combined the primary and secondary task scores into one score. In the funnel task, participants were required to guide three dots on a computer screen towards the target zone (“funnel”). Participants then took a computation span task, which was very similar to the OSPAN, and a dot matrix task. The dot matrix task’s processing component required participants to verify that lines connecting dots on two 3x3 matrixes were equivalent to the lines on the third 3x3 matrix, while its storage component required participants to recall the position of a dot in a 5x5 matrix. Colom et al. (2010) found that WMC was a much stronger predictor of multitask performance than intelligence (correlation of .35 and .22, respectively) and that intelligence did not predict multitask performance when WMC was factored out. Interestingly, while both the processing and the storage component of the WM span tasks were significantly correlated to multitask performance, the processing component had the stronger correlation (0.41 vs. 0.33), suggesting that the processing component should be considered in computing working memory capacity scores. While it is true that the multitasking measures used in Colom et al. (2010) are different from those used in Bühner et al. (2006) and König et al. (2005), Salthouse and Miles (2002) found evidence to suggest that multitasking ability is one construct, and, therefore, as in the case of working memory, that different multitasking tasks can be used interchangeably.

Salthouse and Miles (2002) compared the ability to multitask across three separate multitasking activities. In their experiment, 150 adult participants completed three multitasking activities, each with the same secondary task, but with a different primary task: Directions Primary Task; Series Completion Primary Task; Paired-Associates Primary Task. The secondary
task required participants to keep their mouse cursor on a randomly moving target. The Directions Primary Task, which was given first, involved participants keeping track of directions that they were presented with audibly. Second, the Series Completion Primary Task required participants to logically complete a numerical sequence that was presented to them audibly. Third, the Paired-Associates Primary Task required participants to keep track of the second number of a series of pairs of numbers presented to them. Participants also performed additional cognitive tasks, which included the Letter Series, Spatial Relations, and Paper Folding (for Letter Series Task see Noll & Horn, 1998; for the Spatial Relations Task see Bennett, Seashore & Wesman, 1997; and for the Paper Folding Task see Ekstrom, French, Harman & Dermen, 1976). Performance was found to be moderately correlated across the multitasking tasks, while performance on the multitasking tasks was only weakly correlated to the additional cognitive tasks. Additionally, the authors performed a structural analysis with all tasks and found that the multitasking activities represented a single factor separate from the remaining cognitive tasks. These results, according to Salthouse and Miles (2002), suggest that multitasking ability is an individual difference construct. Salthouse and Miles’ (2002) findings suggest that different tasks can be used to measure multitasking performance, and that, therefore, Colom et al. (2010), despite using the Divided Attention Task and the Funnel Task, instead of the SIMKAP used by previous research on the relationship between WMC and multitasking ability (Bühner et al., 2006; König et al., 2005), can be directly compared to these studies.

Together, the results of Bühner et al. (2006), Colom et al. (2010), Salthouse and Miles (2002), and König et al. (2005) appear to demonstrate the relationship between working memory capacity and multitask performance, thereby providing strong motivation for the inclusion of WMC in the current experiment. First, the multitask measure (SIMKAP) used in both Bühner et
al. (2006) and König et al. (2005), has aspects in common with two of the three conditions (processing for form at the depth of identifying & interpreting) of the main task of the current experiment as they all require participants to take note of a specific detail, a number in the case of the SIMKAP and a form in the case of the current experiment, while comprehending a larger set of details, a filled-out calendar in the case of the SIMKAP and a text in the case of the current experiment. While the SIMKAP is certainly not identical to the multitask element in the current study, Salthouse and Miles’ (2002) finding that different multitask activities measure the same construct suggests that they still may be related. Indeed, Salthouse and Miles’ (2002) findings indicate that even Colom et al. (2010), which used a measurement of multitask performance that was much different than the multitask element in the current study, must also be considered. Together, these findings warrant exploring the possibility that performance for two of the three conditions (processing for form at the depth of identifying and interpreting), to be soon explained in greater detail, may be affected by working memory capacity.

Due to the relationship between working memory and reading comprehension, which is the assessment task of the current experiment, and working memory’s strong correlation to multitask performance, given the multitask nature of the tasks proposed here, it seems more than plausible that working memory capacity could play a role in the current experiment. This is likely to be especially true in the Think-Aloud condition, which will require participants to perform the yet additional task of verbalizing their thoughts.

Concurrent Verbal Reports

In concurrent non-metacognitive verbal reports, also called Think-Alouds, participants verbalize their thoughts as they complete a task. In other words, participants say what they are
thinking as they think it, without being given any specific instructions on how to or what to think. Ericsson and Simon (1993) provide a potential example of what a verbal protocol might look like for a participant solving the math equation “36 times 24”:

(1) “4 times 6,” “24,” “4,” “carry the 2,” “12,” “14,” “144…” (p. xiii).

In the case of this verbalization, the researcher would know the precise strategy that the learner used to solve the math equation, information that might not have been otherwise obtained.

Concurrent verbal reports contrast with retrospective verbal reports in that the latter are collected after the participant has completed a task, as opposed to during the task itself. For example, to collect a retrospective verbal report for a participant which solved the math equation “36 times 24,” the researcher might ask a participant to, as Ericsson and Simon (1993) suggest, start their report with “I first thought of…,” and subsequently ask the participant to detail all his/her thoughts while he/she completed the report. While one might imagine that a decent memory would allow a participant to recall his/her thoughts for this simple math problem, with longer and more complex tasks participants may not remember all of their thoughts, resulting in what is known as veridicality or untruthfulness of the recall due to memory decay. Thus, a major advantage of concurrent verbal reports is that they are completed at the same time as the activity, and therefore do not suffer from veridicality.

According to Ericsson and Simon’s (1993) model of Think-Alouds, there are three types of verbalizations: Type 1, Type 2, and Type 3 verbalizations. Type 1 verbalizations are the easiest to report, as they are non-metacognitive verbalizations that mainly require subjects to say what they naturally think anyway to complete the activity. The Think-Aloud example provided above (1) is an example of a Type 1 verbalization according to the authors. Type 2 verbalizations, while also non-metacognitive, require the participant to recode their thoughts,
which are not necessarily verbal (such as about odor, as the authors suggest), so that they can be verbalized. Type 3 verbalizations, also known as metacognitive verbalizations, require the participant to recode and explain their thought processes. Ericcson and Simon (1993) cite Think-Aloud instructions from Wilson and Schooler (1991), in which participants performed a judgment and decision task, as being likely to elicit type 3 verbalizations:

(2) “analyze why you feel the way you do about each alternative” (p. xviii)

Concurrent Verbal Reports in SLA

Second language acquisition research strives to investigate, among other variables, the strategies or processes that foreign language learners employ to learn as they engage in activities such as reading, writing, and interacting with others. To gather data on these internal strategies or processes, researchers use surveys, interviews, retrospective protocols, reaction times, eye-tracking, and concurrent verbal reports. Of all of these methods, however, concurrent verbal reports, eye-tracking and reaction times offer real time data on the internal processes of learners (Leow & Bowles, 2005). Eye-tracking and reaction times, however, as they are limited to eye movements and time data, arguably only offer superficial and indirect data on internal processes. Only concurrent verbal protocols, where participants think aloud as they complete the task, offer the additional advantage of providing data which is both direct and thorough, offering a window into learners’ attention, awareness, and processing. As noted in Leow and Bowles (2005), concurrent verbal reports have been used extensively in SLA.

In second language acquisition research, concurrent verbal reports have been used to explore learner strategies (e.g., Cohen, 1984, 1994, 2000; Hosenfeld, 1976; Herwig, 2001; Zimmerman, 1994, as cited in Herwig, 2001), as well as to explore the roles of attention,
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

awareness, and processing on learner intake and production (e.g., Hama & Leow, 2010; Leow, Hsieh, & Moreno, 2008; Leow, 1997; Rosa & Leow, 2004).

In Zimmerman (1994, as cited in Herwig, 2001) and Herwig (2001), Think-Aloud protocols allowed researchers to trace participants’ routes of lexical retrieval in a translation task. In Zimmerman (1994, as cited in Herwig, 2001), Think-Aloud protocols gave the researcher insight into how participants processed and organized lexical items in the process of translation. In Herwig (2001), 1 Norwegian and 3 Irish students of Germanic languages first wrote a story in their native language based on a series of pictures provided. Subsequently, participants thought aloud as they translated their story into each of their second languages, which, in the case of the Norwegian were Swedish, German, and Dutch, while in the case of the three Irish students were German, Dutch and Swedish. With the Think-Aloud data, Herwig was able to discover that participants commonly used retrieval strategies that involved multiple languages for a single translation. In one example, a participant trying to translate ‘complains’ into Dutch first sought an English synonym, then a German form, until finally finding an English form that allowed her to access a Dutch item. Without Think-Aloud data, Herwig (2001) would only know that the participant arrived at a Dutch item from the English word ‘complain,’ missing the important data on the strategy that the learner used to arrive at the Dutch word. Given the detail involved in the retrieval of this one lexical item, a retrospective protocol or post-test survey, by themselves, would likely be insufficient in these studies, given that participants are very unlikely to have a perfect memory of their retrieval of each lexical item. These two studies (Herwig, 2001; Zimmerman, 1994) provide justification to the use of concurrent verbal protocols in SLA research, as they show how this research tool allows researchers to gain insight that is not otherwise attainable.
In Leow (1997), Rosa and Leow (2004), and Hama and Leow (2010) concurrent verbal reports gave researchers insight into the level of attention and awareness of participants during task completion. In Leow (1997), 28 beginning students of Spanish thought aloud while completing a crossword designed to encourage noticing of past tense Spanish stem changes. Leow (1997) used the concurrent verbal reports to determine the level of awareness of participants, which he coded as having +/- cognitive change, +/- meta-awareness, and +/- morphological rule. For example, Leow (1997) coded one participant, who said “…well, I guess I made another mistake, corregir is not corregió, it’s corrigió…” (p. 481), during the Think-Aloud, as having experienced a cognitive change, but not having experienced meta-awareness or an understanding of the morphological rule. This is because while the participant indeed found and wrote down the right answer into the crossword, he did not demonstrate an understanding of why the answer was corrigió, as opposed to corregió. On the other hand, another participant said “…12 down, sí, the stem changes e to i, corrigió…” (p. 482), during the Think-Aloud. From this portion of the participant’s Think-Aloud, it is clear that the participant not only was able to come to the correct answer, but also came to the understanding that corrigió was the correct answer because of a stem-change rule that the crossword helped him to discover. Without the Think-Aloud data, these two participants would not have been able to be differentiated in the final analysis, as they wrote down the same final answer. In addition, for this reason, Leow’s (1997) eventual finding, which was that participants demonstrating a level of awareness similar to that of the second participant performed more accurately than those with an awareness similar to that of the first participant, was made possible by the Think-Aloud protocols. Concurrent verbal reports were also an important research tool in measuring awareness in Rosa and Leow (2004), in which it was found that both explicitness in instructions and high levels of awareness...
were conducive to participants’ ability to recognize and produce new exemplars of Spanish contrary to fact conditional sentences. Support for the essentialness of concurrent verbal protocols in measuring internal processes, for example, the construct of awareness, can be found in Hama and Leow (2010).

Hama and Leow (2010), revisiting Williams’ (2005) study, empirically examined learning without awareness with the use of Think-Aloud protocols. In Hama and Leow’s (2010) study, like in the original study, four artificial determiners were used, which indicated whether the noun they modified was near or far, and whether it was animate or inanimate. In the initial training phase, 88 initial participants first performed a Vocabulary Pretraining Task, in which they practiced the determiners and were introduced to the distance, but not the animacy, aspect of their meaning. The experiment was designed to keep participants from becoming aware of the animacy aspect of the artificial determiners’ meaning. After the Vocabulary Pretraining Task, participants completed the training task, in which they listened to a series of sentences with the artificial determiners, repeated the sentences, identified whether the artificial determiner in the sentence was near or far while verbalizing their reasoning, and imagined the situation presented by the sentence in their minds. While continuing to Think-Aloud, participants completed a multiple choice and a production assessment task on the use of the determiners. Participants then completed a questionnaire probing their awareness of the animacy aspect of the artificial determiners. Hama and Leow (2010) used the Think-Aloud data and the offline questionnaire in order to eliminate participants who had become aware of animacy during the experiment. In their results, Hama and Leow (2010) did not find any evidence for learning among the unaware learners. What is of particular interest, is that the questionnaire, while providing evidence of awareness in a number of cases, failed to provide evidence of awareness of the animacy rule for
two participants who were determined to be aware in the Think-Aloud protocols. This is out of a
total of eleven participants who were found to be aware either in the Think-Aloud protocols
and/or in the questionnaire. Without the inclusion of the Think-Aloud protocols, these two aware
participants would have been coded for as unaware, and included in the final analysis. Also of
note is the contrast in findings between this study and the original study (Williams, 2005), who
did not use Think-Aloud protocols, and did find evidence that unaware participants can learn.
Hama and Leow (2010) is evidence, therefore, that the use of Think-Aloud protocols is not only
a useful research tool, as indicated by Leow (1997) and Rosa and Leow (2004), but may be an
essential one, or at least one that cannot reliably be replaced by a questionnaire, for detecting
internal processes during task completion.

Concurrent verbal protocols, therefore, as demonstrated in a wide variety of studies (e.g.,
Cohen, 1994; Herwig, 2001; Leow, 1997), are an important research tool in SLA, and one that
may even be essential in research on internal processes (see Leow, Grey, Marijuan, & Moorman,
2013 for further discussion of concurrent data elicitation procedures).

Think-Alouds: Measuring Depth of Processing

Think-Alouds were also key in three studies previously explored here (see Depth of
Processing in SLA), in which they enabled Qi and Lapkin (2001), Leow, Hsieh, and Moreno
(2008), and Morgan-Short, Heil, Botero-Moriarty, and Ebert (2012) to explore participants’
levels of processing, and, in the case of Leow et al. (2008) and Morgan-Short et al. (2012),
provided crucial evidence that some participants did not process for meaning.

Qi and Lapkin (2001) used Think-Alouds to measure participants’ “quality of noticing,”
or processing levels, as they compared their writing and a reformulated, corrected version. In
their examination of the Think-Aloud protocols, the researchers identified two levels of processing during language-related episodes, 25 in the case of one participant and 16 for the other, in which participants commented on the differences between their original story and the reformulated one. In instances, they found that the participants exhibited a shallow level of processing in which they merely noticed a difference between their original story and the reformulated one, and a deeper level of processing in which they not only noticed a difference, but were able to provide an explanation for the difference. For example, in the following segment of a participant’s Think-Aloud protocol, shown along with the error and the reformulation, the authors coded the participant as having processed at the level of noticing:

7. Form (subject-verb agreement-noticing only)– Su S2: [Original: Some parents they doesn’t want to share theirs wealth/Reformulation: Some parents do not want to share their wealth]. “Here is different: ‘don’t’ and ‘doesn’t’” (p. 291).

This participant was coded at the lower level of noticing because while she did indeed notice that the reformulated version used don’t where he/she used doesn’t, he/she gave no indication that he/she was aware of why this was the case. On the other hand, in the following segment of another participant’s Think-Aloud protocol, the participant processes at a higher level:


Here, the participant not only notices that what he wrote, gun fighting, was different from the reformulation, gunshot, but also identifies the reason, which was that a noun was needed. Qi and Lapkin (2001) observed that these two different processing levels appeared to result in a difference in writing improvement.
Leow et al. (2008) and Morgan-Short et al. (2012) used Think-Alouds to explore participants’ processing levels and eliminate participants who did not follow instructions. In both studies, participants in the processing for meaning and circle form conditions thought aloud as they read an article in Spanish on the Aztecs while simultaneously circling the target form. In their analysis of the Think-Alouds, Leow et al. (2008) found evidence of, and coded for, three levels of processing: “Level 1- merely circling targeted forms; Level 2- providing a report of attending to the target form; and Level 3- interpreting or translating the targeted form” (p. 679).

In addition, Leow et al. (2008) and Morgan-Short et al. (2012) analyzed the Think-Alouds and found evidence that a number of participants were not processing for meaning or backtracked. Therefore, the Think-Aloud data not only allowed for Leow et al. (2008) and Morgan-Short et al. (2012) to identify processing levels of participants, but they also allowed the researchers to eliminate participants who were not completing the tasks as intended by the experiments, therefore avoiding a potential contamination of the results. Both Leow et al. (2008), and Morgan-Short et al. (2012), who used similar coding criteria for processing, compared level of processing and overall comprehension of the text on the Aztecs. Leow et al. (2008) did not find any evidence of a connection between processing and comprehension due to insufficient data, while Morgan-Short et al. (2012) found that higher processing was related to higher levels of comprehension.

In summary, concurrent verbal protocols allowed Qi and Lapkin (2001), Leow et al. (2008), and Morgan-Short et al. (2012) to differentiate between different levels of processing, which, in two of the three cases (Qi & Lapkin, 2001; Morgan-Short et al., 2012) found evidence that suggests a difference in performance. Think-Alouds also allowed Leow et al. (2008), and Morgan-Short et al. (2012) to eliminate participants who were not following instructions.
Concurrent verbal reports, therefore, appear to be an important tool not only for establishing levels of processing, but also serving as an important methodological tool to address the internal validity of the study.

Reactivity

While concurrent verbal protocols have been demonstrated to be a useful research tool, it has been suggested that, as they require the participant to say what he/she is thinking while conducting a task, they could potentially affect task performance, an effect known as reactivity (cf. Bowles 2010 for a review of this issue).

Ericsson and Simon’s (1996) Theoretical Analysis on the Validity of Think-Alouds

Ericsson and Simon (1984, 1996) theorized about the validity of Think-Alouds. According to Ericsson and Simon (1996), Type 1 verbalizations, as they require participants to simply say what they would normally think anyway (see Concurrent Verbal Protocols for a more detailed explanation of verbalization types), are not likely to be reactive. The authors also assert that Type 2 verbalizations, in which a participant recodes his thoughts to make them verbal, do not affect thought processes either, as they do not require the participant to explain his thoughts. However, the authors do cite a then recent study showing that type 2 thinking aloud can influence and change cognitive processes (xx-xxii; Brinkman, 1990). The authors claim, however, that Type 3 verbalizations, which require participants to make connections and inferences about their thoughts, are more likely to alter the thought processes, and increase the
chance that accuracy will be affected, citing several studies consistent with these assertions (Ahlum-Heath & Di Vesta, 1986; Ballstaedt & Mandl, 1984; Stinessen, 1985).

Empirical Studies in SLA

Leow and Morgan-Short (2004), the first SLA study to empirically investigate reactivity, examined the performance of 77 beginning Spanish learners in text comprehension, written production, and recognition of the formal imperative under Think-Aloud and Non-Think-Aloud conditions. The study did not find a significant difference between the Think-Aloud group and Non-Think-Aloud group for the text comprehension test ($p = .142$), the written production test ($p = .427$), or the recognition test ($p = .818$). Interestingly, although differences did not reach a significant level, the Think-Aloud group did consistently better than the Non-Think-Aloud group, performing better on the comprehension task ($M = 10.03$ vs. $M=8.97$), improving more in both the recognition task (improvement of 5.31 vs. 3.51) and the production task (improvement of 1.23 vs. 1.07).

Rossomondo (2007) examined the effect of concurrent verbal reports on comprehension and the incidental acquisition of the Spanish future tense from lexical temporary indicators. One hundred and forty first semester Spanish students were assigned into one of four groups: +TA and +LTI; -TA and +LTI; +TA and –LTI; -TA and –LTI. Participants read one of two passages written in the future tense (+/- LTI), completed a comprehension task, and then either completed a form-recognition task or a cloze passage production task. No significant difference between the silent group and the Think-Aloud group was found for comprehension. With regards to the form recognition and form production tasks, while the groups were relatively small (12 for form recognition and 9 for production), the strikingly large difference in means between the silent and
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

Think-Aloud groups (5.36 vs. 9.5 for form recognition, and 4.03 vs. 8.00 for form production, respectively) was sufficient enough to find the TA condition to significantly enhance learner performance on the Spanish future tense recognition and production tasks. Given the similarities between Leow and Morgan-Short (2004) and Rossomondo (2007), the different results with regards to reactivity in form recognition and production underscore the potential that an individual variable may be playing a role. Indeed, Goo (2010) who examined the relationship between working memory capacity and reactivity, investigated this very possibility. The potential of an individual difference playing a role in the results is also seen in Sachs and Polio (2007).

Sachs and Polio (2007) found some evidence for L2 Think-Alouds negatively affecting performance when compared to a silent group in one of two experiments. In Sachs and Polio (2007), reactivity was examined in a writing correction task in two separate experiments. In the first experiment, fifteen students (11 Korean, 3 Japanese, and 1 Indonesian) in a high-intermediate ESL course took part in the first study with a within-subjects design. Each participant revised three separate compositions they wrote, for one, they were given written corrections of their errors, for another they were given reformulations of their errors, and for the remaining composition they were given reformulations and asked to Think-Aloud during the task. The participants performed significantly better (made more improvements in accuracy on their essay) under the silent reformulation condition than under the reformulation + Think-Aloud condition (p = .047). In the second experiment, which included 54 participants, a non-repeated measures design, more time between the comparison and revision stages, and a control group, but was otherwise similar to the first experiment, thinking-aloud was not found to be reactive (p = .77). If the results in these two studies were only slight, one could reasonably attribute the
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

difference to the two distinct participant comparison methods employed. However, given that the
two results were so different (p = .047 vs. p = .77), it would have been worth exploring the
possibility that individual differences also played a role, especially given the relatively few
number of participants in the first experiment. Given the potentially high cognitive load of the
correcting errors while thinking aloud, working memory capacity could have played a role in
explicating these differences. The possible impact of working memory is underscored by the
finding that working memory capacity has been found to be significantly related to writing
ability and note-taking (Benton, Kraft, Grover & Plake, 1984; Kiewra & Benton, 1988). Sanz,
Lin, Lado, Bowden and Stafford (2009) is another example of a study on reactivity which
includes two experiments yielding different results.

In Sanz, Lin, Lado, Bowden and Stafford (2009), reactivity was investigated in
instructional computerized lessons in two separate experiments. In the first experiment, 24
college participants between the ages of 18 and 35 received a grammar lesson, practice, and
feedback on how noun phrases are assigned semantic functions in Latin (essentially, they learned
how to determine “who was doing what to whom,” p. 46). The participants then took an aural
interpretation test, a grammaticality judgment test, and a sentence production test. In this
experiment, no difference in accuracy was found between the Think-Aloud and the Non-Think-
Aloud groups. In the second experiment, similar to the first except that participants did not
receive the grammar lesson, the Think-Aloud group improved significantly more than the control
group from the pre-test to the post-test (Time x Group) on the grammaticality judgment and
production tests, suggesting positive reactivity (p = .039, p = .035, respectively). The authors
suggest that the finding of reactivity in the second experiment, but not in the first experiment,
was a result of the Think-Aloud condition allowing participants to make form-meaning
connections in absence of the grammar lesson. Nonetheless, if this is indeed the case, it is not clear why the Think-Aloud condition has only significantly improved form score in one out of four SLA studies on reactivity for form accuracy (in Rosomondo, 2007; but not in Bowles and Leow, 2005; Bowles, 2008; Leow & Morgan-Short, 2004). Perhaps it could be argued that this is because Sanz et al. (2009) involved more practice, but that would be counterintuitive, as the Think-Aloud condition’s aide in identifying form-meaning connections would be expected to compensate more in the absence of practice, just as Sanz et al. (2009) assert that it compensated for the lack of a grammar lesson in their experiment. Another possible explanation, given different findings for a number of similar studies (Leow & Morgan-Short, 2004; Rosomondo, 2007), and experiments within studies (Sachs & Polio, 2007; Sanz et al., 2009), is that an individual difference is playing a role. Adding to the potential that an individual difference may be playing a role is the results of a similar but differently focused, experiment conducted by Stafford, Bowden, and Sanz (2012).

Stafford et al. (2012) explored the impact of pre-practice grammatical instruction and explicit feedback on the intake and production of the Latin thematic agent/patient roles, and reported on the impact of a subgroup of participants that thought aloud. In the experiment, 65 Spanish-English bilingual adults in the United States were randomly assigned to four treatment groups, which varied in that two received prepractice grammatical instruction and two received more explicit feedback (+/-GE and +/-EF). Twenty-six participants also verbalized their thoughts. All participants completed task-essential practice on the target item. Participants took four immediate and delayed tests on the target form: (1) written interpretation; (2) aural interpretation; (3) grammaticality judgment; and (4) production. Stafford et al. (2012) found evidence of an immediate impact of the prepractice grammatical instruction in the
grammaticality judgment test and the written production test. They also found that the combined prepractice grammatical explanation and more explicit feedback group scored significantly higher on the delayed written production post test. With regards to reactivity, Stafford et al. (2012) found no evidence of any impact of the Think-Aloud condition on the results of the study. It should be pointed out, however, that, assuming the participants were equally divided amongst the four main conditions, that only between 6 and 7 participants would have been in each group, potentially favoring the non-significant finding. Given the similarity to the previous study explored, Sanz et al. (2009), the absence of reactivity in Stafford et al. (2012) compared to the reactivity found in one of two experiments in Sanz et al. (2009) provides further evidence that an individual difference may be playing a role. One might speculate from the results of Yoshida (2008) that this individual difference may be related to the cognitive resource threshold of participants.

Yoshida (2008) empirically examined the effect of Think-Alouds, outlining, and embedded questions on recall. In the experiment, 64 English majors at a Japanese university were randomly assigned to one of three tasks to perform while reading: (1) outlining; (2) answering embedded questions; and (3) reading-only. The participants were also randomly assigned to Think-Aloud and Non-Think-Aloud conditions (TANTA). Subsequently, participants produced a written recall of the passage both immediately after the reading and one-week after. Yoshida (2008) found that neither task nor TANTA were significant in a two-way ANOVA on immediate or delayed recall. The use of English majors as participants, however, is a limitation to Yoshida (2008), as participants majoring in a language cannot be expected to be equivalent to the general population of language learners. Motivation and language aptitude, two important individual variables to language learning, could reasonably be expected to be higher in such
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

participants. Additionally, given that the study was performed on recall, the extent to which comprehension and not memory is being measured is unclear. Indeed, as Leow et al. (2008) note, “it could be argued that recall tasks might depend too much on the memory capacities of participants, which might compromise the reliability of comprehension results” (pp. 672-673). It might also be argued that the aspect of memory may particularly affect any potential impact of the Think-Aloud condition. Indeed, Yoshida (2008) specifically notes that any potential negative impact of the Think-Aloud condition may have been “cancelled out” by helping participants “comprehend and remember the text by reading aloud difficult parts repeatedly” (p. 206, enhancement added). Certainly, any amount that the Think-Aloud condition does help in terms of memory would be particularly facilitative in a written recall comprehension test, thus making the authors “cancelling out” possibility all the more likely. An interesting finding of the author’s follow-up analysis was that the Think-Aloud condition was near significant in terms of its impact on the performance of the secondary reading tasks. As Yoshida (2008) argues, the cognitive resources of participants may have been overloaded by the reading task and thinking aloud simultaneously, potentially leaving them with insufficient resources to allocate to the secondary tasks.

Reactivity and Working Memory Capacity

Goo (2010) was the first to empirically examine the possibility that working memory capacity may play a role in reactivity. In Goo (2010), 42 participants took two working memory capacity span tasks, read a Spanish text on with 20 instances of the first person plural form of the immediate future and then took a comprehension test and written production test. For the reading comprehension task, a regression analysis showed that WMC was not a significant predictor. As
this contradicts previous findings (Daneman & Carpenter, 1980; Harrington & Sawyer, 1992; Walter, 2004), Goo (2010) cautioned that the non-significant results for Working Memory Capacity may be due to the confounding effect of the Think-Aloud condition. Also for the Reading Comprehension Task, the difference between the Think-Aloud (TA) and the Non-Think-Aloud (NTA) groups (TANTA), measured with a t-test, did not reach statistical significance either. For the written production task, a regression analysis showed WMC predicted the post-test score. In a repeated measures ANCOVA with WMC as a covariate, however, TANTA was not significant. To analyze the relationship between working memory and reactivity, a Mann-Whitney U test was calculated, comparing the scores of the high working memory group to the low working memory group. Contrary to the author’s expectations, under the TA condition, lower working memory capacity individuals performed better, although not significantly so, than higher working memory capacity individuals.

The results of Goo (2010), nonetheless, need to be viewed with caution for a variety of reasons. First, the target form used was the first-person plural of the immediate future. With this target form, the answer was always vamos (“we are going”) plus a (to) plus the infinitive, which is given, not allowing for sufficient variation in participant scores, as participants who learn this basic formula are unlikely to make any mistakes, just as participants who do not are unlikely to fill in any blank correctly. Indeed, in the results of Goo (2010) in the low-WMC-TA condition, while one individual filled in every blank correctly, the remaining four participants in this group did not fill in a single blank correctly, highlighting this very problem. Second, after the elimination of the middle working memory capacity group, at most eight participants are left per cell for the extreme group comparison looking at the effect of working memory capacity on reactivity. This makes statistically significant results very difficult to obtain without a large
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

difference in means and/or very low standard deviations. Given the important limitations of Goo (2010), the question of whether there is a relationship between working memory capacity and reactivity remains inconclusive. Nonetheless, the results of two other SLA studies do highlight the potential that when processing requirements are raised performance may be affected by the Think-Aloud condition (Bowles, 2008; Bowles & Leow, 2005).

Reactivity and Metacognitive Concurrent Verbal Reports

Bowles and Leow (2005) and Bowles (2008) investigated the difference in accuracy between a metacognitive Think-Aloud group (Type 3), a non-metacognitive Think-Aloud group (Type 2), and a control group. In Bowles (2008), reactivity was examined in a problem-solving task where 194 first-semester learners of Spanish navigated through a maze as they made decisions about how to form the Spanish verb *gustar*, which clitic to use with it, and where to place the preposition *a*. The participants in the study were assigned randomly to one of six different experimental groups that each underwent one of three different types of verbalization (metalinguistic Think-Aloud, non-metalinguistic Think-Aloud, and silent) and one of two different types of feedback (implicit or explicit). The results showed that the control group performed significantly better than the metalinguistic group in the production of previously encountered exemplars for both feedback conditions, suggesting reactivity for the metacognitive group. In Bowles and Leow (2005), 45 fifth-semester Spanish learners read a text containing the target form (the pluperfect subjunctive) and subsequently completed a comprehension test and a written production test on the target form. The only statistically significant effect on accuracy was found between the metacognitive group and the non-metacognitive group on text comprehension, with the non-metacognitive group performing significantly better (p < .05).
simplest conclusion from the results of Bowles (2008) and Bowles and Leow (2005) is that the explaining of one’s thoughts out-loud in and of itself is the factor negatively affecting performance. Another possible interpretation, supported by the resource-sharing account of working memory capacity (Daneman & Carpenter, 1980), and Jourdenais’ (2001) suggestion that thinking aloud might constitute an additional task, is that the effect on performance was due to the participants being pushed beyond their available resources by totality of tasks, which included thinking-aloud, explaining one’s thoughts, and navigating through a grammatical decision maze in the case of Bowles (2008), or comprehending a text in the case of Bowles and Leow (2005). That is to say, performance may have been affected in this experiment with the addition of any sufficiently cognitively demanding task to the non-metacognitive Think-Aloud group.

Morgan-Short et al. (2012) explored the possibility of cognitively demanding tasks resulting in reactivity in SLA in their conceptual replication of Leow et al. (2008). Morgan-Short et al., investigated, in addition to other research questions, the influence of thinking-aloud on the ability of participants to process a text for meaning and form simultaneously. Unlike in Leow et al. (2008), Morgan-Short et al. (2012) eliminated those participants in the Think-Aloud condition that reread. In total, Morgan-Short et al. (2012) eliminated 53 participants, or approximately 25% of the original size of the Think-Aloud group, for rereading. Morgan-Short et al. (2012), however, did not eliminate the individuals in the Non-Think-Aloud condition (not present in Leow et al., 2008) that reread. Morgan-Short et al. (2012) did find the Think-Aloud condition to have a significant impact on the results, although they emphasize that the effect size found was too minimal \( n^2 = 0.01 \) to be of practical value, citing Ferguson (2009, p. 533).
The effect size for the Think-Aloud condition, however, must be viewed in light of the potentially confounding factor of rereading. Morgan-Short et al.’s (2012) conclusion with regard to reactivity while participants are processing for form and meaning, therefore, needs to be viewed with caution, motivating the examination of reactivity in the current proposal, which, as in Morgan-Short et al. (2012), looks at the ability of participants to process for form and meaning at the same time.

Purpose of the Study

In the current study, following Leow et al.’s (2008) suggestion to look at the effect of different depths of processing, three depths of processing were established: (1) Processing for meaning only; (2) Processing for meaning and for form at the depth of identifying; (3) Processing for meaning and for form at the depth of interpreting. In the case of Processing at the depth of identifying, the current study used the Spanish aspect, and required participants to click on the past forms regardless of whether they were imperfect or preterit forms. This depth of processing is based on Leow et al.’s (2008) coding of Levels I (Circling only) and II (Circling and reporting), as participants assigned to this group may also report attending to the form in the concurrent verbal report. This depth of processing for form is low as it only required the participants to process for a basic facet of the verb conjugation that has a strong connection to meaning, that is, that the verbs indicate past tense. For Processing at the depth of interpreting, participants were instructed to identify (i.e., find) and interpret the aspect of each of the past forms during the reading, as either the imperfect or the preterit by clicking once or twice, respectively, on the verbs. Interpreting the aspect of the target form as either imperfect or preterit, as opposed to simply identifying the past forms, required the forms to be processed more
deeply as the aspect is beyond that which is essential for a basic understanding of the meaning of
the conjugated verb. This depth of processing follows Leow et al.’s (2008) coding of Level III
(interpreting or translating the target form), with the important distinction that the current study
only codes processing at the depth of interpreting when the form has been interpreted. The
connection between Leow et al.’s (2008) coding of Level III processing and the current study’s
coding of processing at the depth of interpreting is most clear in the case of the former’s coding
of participants processing of the morpheme –n (plural marker of the verb in Spanish). Leow et al.
(2008) counted those participants that said aloud “they” and the verb in English in response to a
Spanish verb in the plural form as having processed the –n morpheme at the third depth of
processing. Whereas Leow et al. (2008) specifically labeled this as an instance of translation,
presumably as opposed to interpreting given the study’s inclusion of both in their
operationalization of Level III, the current study, while in agreement about the depth of
processing, holds that it can be classified as this depth of processing because it is an instance of
interpreting through translation. This is because even if a participant mistranslates a part of the
phrase, such as the tense, the participant still shows evidence of interpreting –n as being a marker
of the plural if he/she uses “they” and the verb in their translation. Labeling a past form’s aspect
is similar to translating the morpheme –n as “they” in that in both instances the participant
processes the function of the target forms.

In addition, the current thesis explored the impact of amount of processing (i.e., the
number of verbs that were identified or interpreted). While participants that identified less than
60% of the forms were eliminated in this study, as in previous research (Greenslade et al., 1999;
Leow et al., 2008; Morgan Short et al., 2012; VanPatten, 1990; Wong, 2001), this still leaves
substantial variation in terms of the actual number of forms that participants process. Given that
VanPatten’s (2004) Primacy of Meaning Principle implies that processing for form detracts from comprehension, processing more for form has the potential of being negatively correlated to comprehension.

The current dissertation also sought to address the methodological concerns of Leow et al. (2008) and Morgan-Short et al. (2012). First, in order to limit participants’ access to additional cognitive resources, two countdown timers and video instructions were employed. Second, in order to limit the potential of mediating variables playing a role on the results, depths of processing were randomly assigned, as opposed to being coded for at the conclusion of the experiment. Third, due to a methodological limitation in Morgan-Short et al.’s design, specifically allowing for rereading in the NTA group but not in the TA group, a NTA group was again used here to examine reactivity together with an electronic design that both disallowed backtracking, as suggested in Leow et al. (2008), and allowed for rereading to be monitored with a screen recording. Moreover, as Morgan-Short et al. (2012) recommended, participants were specifically instructed not to reread.

The relationship between working memory capacity and reading comprehension found in previous research (e.g., Harrington & Sawyer, 1992; Walter, 2004) was again explored in the current study. In light of the findings of these studies (e.g., Harrington & Sawyer, 1992; Walter, 2004), as well as the relationship between reading comprehension and multitask performance (Bühner, König, Pick and Krumm, 2006; Colom et al., 2010; König et al., 2005), the possibility that working memory capacity is a mediator variable was explored in research questions VI and VII. These research questions examined the relationship between working memory capacity and depth of processing for form, as well as the relationship between working memory capacity and reactivity.
Research Questions

I. Does Depth of Processing for form (control vs. identifying vs. interpreting) while processing an L2 text for meaning have an impact on comprehension?

II. Does Reactivity (TANTA) have an impact on L2 comprehension?

III. Is there any interaction between Depth of Processing for form and Reactivity (TANTA) with regards to their impact on L2 comprehension score?

IV. Is there a relationship between amount (number of clicks) of processing for form and L2 comprehension score?

V. Is Working Memory Capacity related to L2 comprehension?

VI. Is there any interaction between Working Memory Capacity and Depth of Processing for form in their relationship to L2 comprehension score?

VII. Is there any interaction between Reactivity (TANTA) and Working Memory Capacity in their relationship to L2 comprehension score?
CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

Introduction
This chapter begins with the Pilot Study, including a discussion of its results, and ends with the full research study.

Pilot Study
The pilot study was conducted to make an assessment as to an appropriate time pressure for the Reading Task (below), to test the materials, and to probe the likelihood of obtaining significant results in the full study.

Participants
Nineteen college-age students enrolled in Intro II or Intermediate I Spanish at Georgetown University were recruited for the original pool of the pilot. Sixteen of these participants were randomly stratified by working memory capacity to one of three Depth of Processing Levels. In order to determine an appropriate time pressure for participants in the full study, in the pilot, participants were also randomly stratified by Depth of Processing Level to two different time pressures to read the experimental texts: (1) three minutes and thirty seconds and (2) four minutes and thirty seconds. Of the three remaining participants, one signed up late for the study, and, due to scheduling restraints, participated in session two before session one, and thus was unable to be randomly stratified by WMC. The remaining two participants, one of which had a high WMC score, and one who had very low WMC, were used to explore the amount of time that a participant with no time-pressure would take in the Think-Aloud, processing for meaning
only group. As these three participants were placed into groups, as opposed to being randomly stratified into groups, while their results would be briefly referenced in the results section, they were not part of the main statistical analyses. Of the sixteen stratified participants, one participant was eliminated for not showing evidence of processing for meaning. In addition, three participants clicked on less than the minimum percentage of target forms identified by VanPatten (1990) as demonstrating processing for form (60%). Of these three participants, one also did not correctly interpret a minimum of 75% of the past forms (he/she interpreted 72% correctly). While participants who fail to reach either of these two minimum percentages will be eliminated in the final analysis of the full study in order to allow for a fair comparison to previous studies (Greenslade et al., 1999; Leow, Hsieh, & Moreno, 2008; Morgan-Short et al, 2012; VanPatten, 1990; Wong, 2001), they were maintained in the analysis of the results of this pilot. One participant also noted advanced knowledge of another Romance language (Latin), while another provided unsolicited information on the survey that he had a documented learning disability.

Linguistic Target
The linguistic target of the pilot was the Spanish past aspect. In Spanish, as in many other Romance languages (e.g., French, Portuguese), the past is expressed with either the imperfect or the preterit tense, based on the meaning and the context. With the exception of three irregular verbs (ser “to be”, ver, “to see”, ir, “to go”), the imperfect aspect follows a predictable pattern, ending in -aba, -abas, aba, ábamos, abais, aban for verbs whose infinitives end in -ar, and in -ía, -ias, ia, íamos, íais, ían for verbs whose infinitive ends in –er or -ir. The preterit aspect is very irregular, but its form can still be clearly distinguished from the imperfect as it never ends in –ía.
or aba. The Spanish past aspect is ideal for the current study because it neatly allows for two depths of processing: Identifying, meaning that the participant identifies that the verb is in the past; and Interpreting, meaning that the participant not only notes that the verb is in the past tense but makes a mental note of which aspect of the past tense the verb is in.

Experimental Texts

Two reading passages were used in the pilot, *Frida Kahlo* and *Los Aztecas* (‘The Aztecs’). The first of these two reading passages, *Frida Kahlo*, served as a practice exercise to ensure students understood the activity.

The Practice Reading Passage, *Frida Kahlo*, was a modified version of an authentic biography on Frida Kahlo found on the cultural section of the family-oriented website *Padres Hispanos* (‘Hispanic Parents’). The text was shortened to 171-words long in order to reduce time spent on the practice activity. The passage was also modified to equally distribute the aspect of the past tense verbs between the imperfect and the preterit tenses. In total, the text included 14 target forms, 7 in the preterit tense and 7 in the imperfect (See Appendix A).

The Main Reading Passage, *Los Aztecas*, was taken from an online cultural website and first modified by Leow et al. (2008), with additional modifications being made for the pilot. Los Aztecas is 373 words long and is about the ancient Aztec civilization and their much-revered Sun God, Tonatiuh. As in the case of the practice passage, *Los Aztecas* was modified to equally distribute the new target forms, the imperfect and the preterit. Additional minor changes were also made to allow for these additional exemplars of the target form to be semantically and contextually appropriate (Appendix B).

Both texts were computer-delivered in order to control for time and backtracking and to monitor for rereading. Participants in Level 2 were instructed to click on verbs in a past tense,
and participants in Level 3 were required to click once for imperfect verbs and twice for preterit verbs (cf. Procedure).

Materials

Think-Aloud Tasks. The participants in the Think-Aloud group were told to think aloud, with written instructions adapted for this study from Goo (2010, pp. 726-727). These instructions asked participants to think aloud during the entire experiment without planning or explaining their thoughts. Participants then performed a Think-Aloud practice task in which they were asked to verbalize their thoughts while they calculated the total cost of a grocery list. Participants were also given reminders, when necessary, to continue thinking-aloud.

Assessment Materials

Brief Surveys. Participants took two brief surveys, one at the beginning and one at the end of the experiment. The Pre-Test Survey (Appendix C) asked participants whether they were taking an Intro II or Intermediate I Spanish class, how old they were, what languages they were able to speak or spoke at different age intervals, and whether or not they have any non-corrected hearing or visual impairments. The Post-Test Survey (Appendix D) asked participants what percentage of the content of Los Aztecas article they knew before reading the article, and what percentage helped them answer the Multiple-Choice Comprehension questions. Participants were also asked to rank their effort on a scale of one to five. An e-mail was subsequently sent to participants asking them how well they felt they were able to understand the meaning of the passage on a scale of 1 to 5 and to explain their choice in 2-3 sentences in English.
**Operation Span Task.** Goo’s (2010) version of the OSPAN task, “originally developed by Turner and Engle (1989)” (Goo, 2010, p. 724), was used in this study, with the addition of a 4 minute 10 second instructional video developed by the current researcher. In the instructional video, participants heard an explanation of the task, which visually includes Goo’s (2010) written PowerPoint instructions, and then saw an example of the researcher taking the practice Operation SPAN task. After watching this video, the researcher asked participants if they had any questions, and subsequently allowed them to proceed to the practice OSPAN task. After taking the practice task, which included six mathematical operation-word pairs divided evenly into three sets, participants were asked if they had any remaining questions and were allowed to proceed to the actual, to-be-analyzed, OSPAN task. The actual task included forty-two mathematical operation-word pair slides, each of which displayed one mathematical operation (e.g., 2+8 =10) followed by a monosyllabic English word (e.g., BANK). The participants were instructed to read aloud each mathematical operation-word pair. Following each operation-word pair slide, another slide was presented asking the participant to mark on their answer sheet whether the mathematical equation was correct or not. The number of operation pairs randomly varied from 2 to 5 in each set. After each set, a slide prompted the participants to recall the words presented after each mathematical operation in the set. They were “given 8 seconds for set size 2, 11 for set size 3, 14 for set size 4, and 17 for set size 5” (Goo, 2010, p. 725).

**Multiple-Choice Comprehension Tasks.** Participants took two Multiple-Choice Comprehension Tasks: The Multiple-Choice Comprehension Practice Task and The Multiple-Choice Comprehension Task.

The Multiple-Choice Comprehension Practice Task consisted of four questions based on the content of the Practice Reading Passage, *Frida Kahlo* (See Appendix E).
The Multiple-Choice Comprehension Task was a modified and expanded version of the Multiple-Choice Task in Leow et al. (2008), which contained multiple-choice questions based on the content of the Main Reading Passage, *Los Aztecas*. Modifications were made to ensure that the necessary changes made to the text were reflected in the questions, and that the number of multiple-choice questions per paragraph corresponded to the number of target forms. For example, a paragraph with two imperfect and two preterit forms (or four total past forms) also contained the information necessary to answer four comprehension questions. This doubled the total number of multiple-choice questions included in Leow et al. (2008) to 20 (*Appendix F*).

Both the Multiple-Choice Comprehension Practice Task and the Multiple-Choice Comprehension Task were administered electronically to maintain a consistent electronic format throughout the second session.

Procedure

In the first session, a survey and an operation working memory span task were administered to participants. The operation working memory span task was preceded by an explanatory video on this task and three practice sessions that were not part of the final data analysis. The same operation span task that Goo (2010) used was employed in this study.

Based on the participants’ working memory capacity, participants were temporarily assigned into three levels of working memory capacity (high, medium, low) by z-score: high working memory (6), for z-scores above .5; medium working memory (5), for z-scores between -.5 and .5; and low working memory (5), for z-scores under -.5 (following Mackey et al., 2002; Goo, 2010). Prior to the second session, 16 of the participants in each of the three working
memory capacity groups were randomly stratified to three Depth of Processing levels such that each group was balanced in terms of working memory capacity:

1. Level 1: Control or Reading for meaning only (5)
2. Level 2: Reading for meaning and processing forms at the level of identifying (5)
3. Level 3: Reading for meaning and processing forms at the level of interpreting (6)

Participants were subsequently randomly stratified to the Think-Aloud and Non-Think-Aloud groups by Depth of Processing Level.

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>NTA</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Participants were also randomly stratified by Depth of Processing Level to two different time-pressures: (1) three-minutes and thirty seconds; (2) four-minutes and thirty seconds. The time-pressure consisted of a countdown timer at the top of the screen. It was different from a time-limit in that, unbeknownst to the participants until the timer reached zero, a pop-up box appeared asking them if they had finished reading, and if they had not they were allowed to finish reading. The purpose of the time-pressure was to encourage participants to complete the task as quickly as possible, and not take additional time to compensate for the additional processing requirement of Depth of Processing Levels 2 and 3.
In the second session, which took place 2 weeks after the first session\textsuperscript{vi}, all participants were screen-recorded, and participants in the TA group were also voice-recorded. Participants in the TA group were instructed to think aloud throughout the experiment. First, participants in both the TA group and the NTA group watched and listened to video instructions that explained to them the tasks of this session, which emphasized the importance of paying attention to meaning. The TA group then read instructions on thinking aloud and completed the mathematical Think-Aloud practice task (see above). All participants subsequently read the modified version of the practice passage \textit{Frida Kahlo} on the computer. Participants in Depth of Processing Levels 2 and 3 had an additional task to complete while they read. Participants in Level 2 were required to click on past forms in the passage in order to minimally induce processing at the depth of identifying. In order to induce processing at the depth of interpreting, participants in Level 3 were required to interpret each verb in the past form as either the preterit or the imperfect by clicking once (imperfect) or twice (preterit). In order to enable clicking, computer software for the experiment was developed by Gregory S. Mercer, Senior Software Engineer at the Stanford Graduate School of Business and the researcher’s father. The software highlighted and recorded the words participants clicked on. In the case of condition 3, ‘[IMP]’ appeared after the word following one click, and ‘[PRET]’ following two clicks. When time ran out, a pop-up screen appeared asking participants whether or not they had finished the reading. Participants that indicated that they had not finished the reading at this point were redirected to the reading and allowed to finish. Otherwise, they proceeded to the Multiple-Choice Practice Comprehension Test. Participants were not allowed to return to the text once they begun the Multiple-Choice Comprehension Practice Task. After completing the Multiple-Choice Practice Task, participants read the Aztecs passage, and subsequently took the Multiple-Choice Comprehension Test, under
the same conditions that they performed the practice tasks. At the conclusion of the experiment, participants took the Post-Test Survey. The one participant that signed up late performed the same tasks as the Depth of Processing Level 1, Think-Aloud participants, and was given a time pressure of three-minutes and thirty seconds. The remaining two participants that were not stratified by working memory capacity performed the same tasks as Level one, Think-Aloud participants, but did not have a time-pressure.

Elimination Criteria

In the full study, participants in Levels 2 and 3 were excluded from the analysis if they did not process at their assigned level of processing. Participants in Level 2 were excluded if they do not process minimally at the depth of identifying. Similar to previous studies (e.g., Leow et al., 2008; Morgan-Short et al., 2012; VanPatten, 1990), identifying (referred to as noticing/circling in Leow et al., 2008) was operationalized as clicking on at least 60% (12 out of 20) of the target forms. Participants in Level 3 were excluded if they did not process at the depth of interpreting. Processing at the depth of interpreting was operationalized with two criteria. First, participants needed to minimally label 60% of the target forms. Second, in order to control for guessing, which would only involve identifying the target form, participants must have correctly interpreted at least 75% of the forms they identify. Seventy-five percent was chosen as the threshold because participants who identify 80% of the target forms (mid-way between the 60% minimum and the 100% maximum) have less than a 5% chance of correctly guessing 75% or more of the forms they interpretedvi. In the pilot, three out of a total of eleven participants in Level 2 and 3 did not click on the minimum percentage of past verb forms (Level 2 participant: 45%; Level 3 participants: 40%, 55%), while one out of six Level 3 participants did not meet the
75% minimum for correctly interpreting forms. While participants doing so were eliminated in the full study in order to allow for a fair comparison to previous studies (e.g., Leow et al., 2008; Morgan-Short et al., 2012), data from these participants was included in the analysis in the current pilot, as they still did follow the basic instructions of their levels, clicking on past forms in the case of Level 2 and clicking on and labeling past forms in the case of Level 3.

As in Leow et al. (2008), processing for meaning was operationalized as “any sustained effort to read the text primarily for meaning” (p. 678). For the Think-Aloud group, the concurrent verbal reports, total time data, the Practice Multiple-Choice Comprehension Task, and answers in the Post-Test Survey were analyzed to determine if participants were processing for meaning. With regards to the Non-Think-Aloud group, as the concurrent verbal reports were not available, the screen recording, total time data, the Practice Multiple-Choice Comprehension Task, and answers in the Post-Test Survey were analyzed to make this determination. Out of a total of 16 stratified participants, one participant was eliminated for not processing for meaning.

Participants that did not finish the story when the time-pressure was up were not eliminated. This was because the purpose of the time-pressure was not to ensure that all participants finished the tasks in the same or approximately the same amount of time, but, rather, to encourage participants to complete the tasks as quickly as possible.

Participants that rated their knowledge of the material covered in the Aztecs passage as 50% or more in the Post-Test Survey would have been considered for elimination. All participants rated their knowledge between 0 and 20%, and, therefore, no eliminations were necessary for this reason.
Scoring

**Multiple-Choice Comprehension Tasks.** One point was awarded per correct answer for a maximum total of four points for the Multiple-Choice Practice Comprehension Task and twenty points for the Multiple-Choice Comprehension Task.

**Working Memory Capacity.** The Working Memory Span Scores were calculated by combining the processing and recall portions of the Working Memory Span Task. The processing part of the working memory tasks was scored by giving one point to each correct answer and computing a percentage for each working memory span task. The recall portion of the working memory tasks was scored following the partial-credit load scoring procedure, as in Conway et al. (2005) and Goo (2010), dividing the number of recalled items by the total number of items (following Conway et al., 2005; Goo, 2010). Working Memory Score was then calculated by averaging the percentage recall score with the percentage processing score on the OSPAN.

Quantitative Analysis

As participants were recruited from two different levels of Spanish, Intro II and Intermediate I, a t-test was first performed in order to determine whether or not there was a significant difference between these two groups with regards to their scores on the Multiple-Choice Comprehension Task. No significant difference was found between the Intro II and Intermediate I groups, and, therefore, they were combined in subsequent analyses of the data. See Table 2 for descriptive statistics.
In addition, as participants reported different levels of effort exerted in the completion of the tasks, a biserial correlation was performed between participants’ scores on the Multiple Choice Comprehension Task (MCCT) and the levels of effort reported on the post-test survey. In the analysis, no significant correlation was found between effort and MCCT scores, $r(15) = -.354$. Interestingly, the value for $r$ indicated that participants with higher reported levels of effort actually scored lower than those with lower levels of reported effort. Given that this result was not significant, however, participants with different levels of reported effort were combined for further analyses.

The amount of time that the participants took to complete the reading varied widely, ranging from 134 to 469 seconds. Participants in Depth of Processing Level 1 ($M = 246.60$, $SD = 126.06$) took less time than participants in Level 2 ($M = 283.50$, $SD = 118.38$) and 3 ($M = 271.17$, $SD = 89.32$), however, Depth of Processing was not found to be a significant factor in the one-way ANOVA performed with Time in Seconds as the dependent variable, $F(2, 12) = .135$, $p = .875$. In addition, in an independent samples t-test, Time Pressure was found to not have a significant impact on Time Seconds, $t(13) = .159$, $p = .434$.

The first, second, and third research questions sought to examine the impact of Depth of Processing, TANTA, and the interaction between Depth of Processing Level and TANTA on
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

comprehension. In order to address these research questions, a 2-way ANOVA was performed with Depth of Processing and TANTA as independent factors, and MCC as the dependent factor. A main effect was found for Depth of Processing, $F(2,9)= 7.482$, $p = .012$, but not for TANTA, $F(1,9) = .005$, $p = .944$, or interaction between Depth of Processing and TANTA, $F(2,9) = .287$, $p = .757$. Table 3 includes descriptive statistics for the data.

Table 3

<table>
<thead>
<tr>
<th>Depth</th>
<th>TA</th>
<th>NTA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>($n = 8$)</td>
<td>($n = 7$)</td>
</tr>
<tr>
<td>1 (Control)</td>
<td>12.50</td>
<td>11.67</td>
</tr>
<tr>
<td></td>
<td>(.71)</td>
<td>(.58)</td>
</tr>
<tr>
<td>2 (Identifying)</td>
<td>8.33</td>
<td>9.00</td>
</tr>
<tr>
<td></td>
<td>(1.53)</td>
<td></td>
</tr>
<tr>
<td>3 (Interpreting)</td>
<td>9.33</td>
<td>9.33</td>
</tr>
<tr>
<td></td>
<td>(2.31)</td>
<td>(.58)</td>
</tr>
<tr>
<td>Total</td>
<td>9.75</td>
<td>10.29</td>
</tr>
<tr>
<td></td>
<td>(2.31)</td>
<td>(1.38)</td>
</tr>
</tbody>
</table>

In the post-hoc Scheffé analysis, a significant difference was found between Depth of Processing Levels 1 and 2, and between Levels 1 and 3, but not between Levels 2 and 3.

The fifth’, sixth, and seventh research questions sought to address whether there was a relationship between working memory capacity and comprehension, an interaction between working memory capacity and Depth of Processing in their relationship to comprehension, and an interaction between working memory capacity and TANTA in their relationship to comprehension. In order to do so, a multiple regression analysis was performed on the data with working memory capacity, TANTA, Depth of Processing, working memory capacity * Processing Level, and working memory capacity * TANTA as independent variables, and MCC as the dependent variable. Working memory capacity and TANTA were both not found to be significant predictors. The interaction between working memory capacity and Processing Level
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

and the interaction between working memory capacity and TANTA were also not found to be significant. Depth of Processing, however, was found to be a significant predictor of comprehension. Table 4 summarizes the values for the multiple regression analysis.

Table 4

<table>
<thead>
<tr>
<th>Overall Descriptive Statistics</th>
<th>B</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP</td>
<td>-1.353</td>
<td>-2.471</td>
<td>.035</td>
</tr>
<tr>
<td>TANTA</td>
<td>4.264</td>
<td>.393</td>
<td>.704</td>
</tr>
<tr>
<td>WMC</td>
<td>-15.858</td>
<td>-.907</td>
<td>.388</td>
</tr>
<tr>
<td>DP * WMC</td>
<td>6.634</td>
<td>1.007</td>
<td>.340</td>
</tr>
<tr>
<td>TANTA * WMC</td>
<td>4.264</td>
<td>.393</td>
<td>.704</td>
</tr>
</tbody>
</table>

The regression equation for Multiple Choice Comprehension was: 

\[
MCC = 13.091 - 1.353PC \times .552TANTA - 15.858WMC + 6.634WMC*DP + 4.264WMC*TANTA, \quad R^2 = .437.
\]

The participant that signed up late scored a 7 on the Multiple-Choice Comprehension Test and took 2 minutes and 34 seconds to complete the Aztecs reading.

An analysis of the data of the remaining two non-stratified participants revealed that the very low working memory capacity participant scored a 9 out of 20 on the Multiple Choice Comprehension Test, and took 5 minutes and 53 seconds to complete the Aztecs reading, while the higher working memory capacity participant scored a 14 and took 3 minutes and 44 seconds to complete the Aztecs reading.

Qualitative Analysis

Think-Aloud Data. The Think-Aloud protocols were used in addition to the practice test score, total time data, screen-recordings, and the Post-Test Survey to determine whether or not participants were processing for meaning. The Think-Alouds, screen recordings, and survey data
were found to provide the most direct evidence of participants processing for or not processing for meaning, as no one spent so little time to be coded as definitively not processing for meaning, and seven participants who scored low (2 or less out of 4) on the multiple-choice practice test nonetheless demonstrated in the Think-Aloud protocols, the screen recordings, or the survey data that they were processing for meaning.

Out of the eight Think-Aloud participants, strong evidence of processing for meaning was found in the Think-Aloud data in the form of translating, commenting, and/or hypothesizing about the content of the passage for seven participants. For the remaining Think-Aloud participant\(^{xi}\), however, the Think-Aloud protocol provided minor, insufficient evidence of processing for meaning, in the form of pausing and reading aloud the content of the passage in Spanish, with the exception of reading the dates in English. Evidence that this participant was processing for meaning, however, was found in his/her rating and response in the Post-Test Survey.

In addition, two participants, one of whom was reading the text entirely in Spanish, and the other who was reading the text in Spanish with the exception of translating three words, were reminded to not simply read the text aloud, but to remember to comment. The participants appeared to understand at this point, as they no longer directly read the text aloud.

**Post-Test Survey.** In the Post-Test Survey, participants were asked to rate how well they were able to understand the meaning on a Likert scale of 1 to 5 (see Appendix D), and provide a brief explanation for their rating. Ratings ranged from a 1 (one participant) to a 4 (one participant), with most participants rating their ability to understand the meaning as a 2 (two participants) or a 3 (twelve participants). A rating of a 3 or higher was considered evidence of processing for meaning. The participant whose Think-Aloud protocols showed insufficient
evidence of processing for meaning selected a 3 out of 5, which the participant explained as being low only because he had difficulty recalling the information that he had just read when answering the questions of the Multiple-Choice Comprehension Test. For the participants giving themselves a rating of 2, evidence of processing for meaning was found in the Think-Aloud data. The participant who selected a rating of one was eliminated due to lack of evidence for processing for meaning in the screen recording, as the participant as well as their explanation for their rating (follows).

P1: My comprehension of the passage was very low due to a combination of factors. It was difficult for me to focus on comprehension and identifying the past tense verbs at the same time. There was also a lot of vocabulary that I was unfamiliar with. It was also very distracting to have the other participants in the study talking around me while I was reading. I am certain that I did extremely poorly on the comprehension questions.

In the screen recording, the participant appeared to be scanning for the target items and not focused on processing for meaning in the screen recording. An example of a participant whose response provided strong evidence that they were processing for meaning follows.

P2: I feel like I was able to understand the general message of the reading but due to outside distractions and a little lack of motivation, I didn't entirely catch all of the little details in the article that were crucial to knowing the answers to the following questions. Also, I haven't been practicing my Spanish as consistently as I should have been, therefore my reading comprehension isn't at its optimal level right now.
While this response does indicate that the participant did not capture every detail of the passage, it also makes it clear that the participant was processing for meaning, as he/she indicates that he/she was able to parse out the main ideas of the passage.

Three participants commented in their responses with regards to the difficulty of processing for form and meaning simultaneously. These three participants asserted that the additional task of processing for form made it difficult for them to process for meaning.

P3: I think I would have understood it better if I wasn't focused on finding the imperfect/preterit verbs as well.

P4: I thought that the beginning of the passage was easier to read but I think I got caught up in the preterit imperfect part towards the end.

P5: It was difficult for me to focus on comprehension and identifying the past tense verbs at the same time\(^{\text{xii}}\).

Additional information gleamed from the responses was that one participant specifically mentioned the countdown timer as a factor motivating them to proceed more quickly, and another felt that thinking aloud was distracting.

P6: I was able to understand the majority of the sentence most of the time but if I ran into a word I did not recognize it confused me. The timer was also a factor in which sentences I read more quickly than others.

P7: Also, when I was thinking aloud it felt more like a distraction then helpful.
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

Discussion

The purpose of the pilot was to test reliability and validity of the materials being used and provide insight into how they might be improved upon and revised for the full experiment. As explored earlier in this dissertation, a potential methodological limitation of two of the previous studies examining the impact of dual processing on comprehension is that they did not include a time-limit (Leow et al., 2008; Morgan-Short et al., 2012), and therefore, did not control for the possibility that participants processing for both form and meaning might compensate for the additional burden on their cognitive resources by taking additional time. Seeking to address this issue, participants in the current pilot had a perceived time-limit (i.e., time pressure) in the form of a countdown timer on the screen in order to encourage participants to complete the readings as quickly as possible. The results of the pilot study provide at least initial evidence that the time pressure employed was successful in this regard, as the average amount of time spent reading the main passage ranged between 4 minutes and 6 seconds for Level 1 and 4 minutes and 43 seconds for Level 2, a much narrower gap than that found between the control Level (4.2 minutes) and the most comparable processing for meaning and form condition in Leow et al. (2008), namely, processing for meaning and the third person plural form –n (5.55 minutes).

Two perceived time limits, one starting at 4 minutes and 30 seconds, and the other starting at 3 minutes and 30 seconds were used in the pilot in order to determine which would be most appropriate for the full study. Given the previously mentioned range of times by Depth of Processing (4 minutes and 6 seconds to 4 minutes and 43 seconds), 4 minutes and 30 seconds appears to be the more appropriate perceived time limit. However, it should be noted that 3 participants took longer than 6 minutes, and, therefore, could potentially be viewed as not following the instructions to proceed as quickly as possible.
Another potential concern that was addressed in this pilot was whether or not the task involved in interpreting forms in Level 3 would be too confusing for participants. The possibility was that participants might forget how many times to click on preterit vs. imperfect verbs, or perhaps get so involved in interpreting verbs that they would forget to process the text for meaning. Neither of these appeared to be an issue, as, with regards to the former, only one participant failed to correctly interpret 75% or more of the verbs they selected correctly as either preterit or imperfect, and, with regards to the latter, all participants in Level 3 showed evidence of processing for meaning. Therefore, no change appears to be necessary with regards to the operationalization of Processing Level 3.

While this was not initially a concern of the pilot study, an analysis of the Think-Aloud data revealed that two participants did not think aloud, but, rather, simply read aloud the text. An additional two needed to be reminded to not simply read the entire text aloud. While it is certainly natural to read aloud portions of the text while thinking aloud, and this was quite common, doing so for the entire text indicates that the participant may not have understood the instructions to think aloud his or her thoughts. Alternatively, it may have been due to the modality of the practice task, a mathematical problem, which may have been more natural for the Think-Aloud process, and thus created an unexpected difficulty when the actual task was reading. To ensure that all participants understood what they are expected to do, in the main study the instructions were made clearer so that they would be understood by participants, and they included an example of a Think-Aloud, which was accompanied by a practice task that was relevant to the actual task of the study.

A secondary, but also important, objective of the pilot was to probe for the likelihood of obtaining significant results in the full study. While most of the results of the pilot were not
significant, a significant result was found for the first research question, namely, *Does Depth of Processing for form (control vs. identifying vs. interpreting) while processing an L2 text for meaning have an impact on comprehension?*, with a higher processing level leading to lower scores. Specifically, participants in Levels 2, *processing for meaning and form at the level of identifying*, and 3, *processing for meaning and form at the level of interpreting*, were found to have significantly lower performance compared to participants in Level 1, *processing for meaning only*. Following VanPatten’s (1990) Primacy of Meaning Principle, this potentially may be due to the increased cognitive load of processing for form having an effect on processing for meaning.

With regard to the five remaining research questions of the current pilot, no significant findings were found for Think-Aloud vs. Non-Think-Aloud (TANTA), Working Memory Capacity (WMC) or any interactions between these variables or between these variables and Depth of Processing. Depth of Processing was again found to be significant, this time shown to have a significant *relationship* to comprehension.

Each of the findings of the current pilot, non-significant and significant, need to be interpreted with caution given the small number of participants involved and the large number of variables measured. Another important limitation of the current pilot is the inclusion of participants who did not meet the minimum standards of their respective processing levels. A full study was clearly warranted to further investigate the findings of the current pilot.
Minor Changes

Discussion with committee members, reflection on part of the researcher, and an analysis of the results of the Pilot Study revealed several improvements that the researcher implemented for the full study. These include:

(1) *Including video TA instructions with an example of the researcher thinking aloud.* The remaining instructions in the pilot, with the exception of follow-up instructions within the tasks, were on video, and, therefore, creating video Think-Aloud instructions added to the consistency and fluency of the overall experiment. In addition, it was noticed by the researcher that the video instructions for the remaining aspects of the tasks in the pilot study were particularly effective, as participants demonstrated that they understood them while they completed the tasks. One might point out that unlike written instructions, participants cannot skip video instructions, and are much less able to ignore them. They are also arguably less boring and capture the attention of the participants to a greater degree. One specific aspect of the video instructions for the working memory capacity task was that they included an example of the researcher taking the working memory capacity task. This was one of the motivations for the inclusion of an example of the researcher thinking aloud in the full study. Another motivation for the inclusion of an example of the researcher thinking aloud is that two participants simply read the text aloud instead of thinking aloud, and two additional participants needed to be reminded to not just read the text aloud. The addition of an example of the researcher thinking aloud made it clear to participants what was expected from the Think-Alouds.
(2) *Use a new TA practice task.* In the Think-Aloud practice task of the pilot, participants were asked to verbalize their thoughts while calculating the total cost of a grocery list although the mathematical nature of this task was not consistent with the reading task that the participants were subsequently asked to perform. In addition, it did not afford the researcher the opportunity to correct participants who were just reading aloud as opposed to thinking aloud, as it is not possible to solve a mathematical task by reading aloud. Therefore, a new Think-Aloud practice task, that reflects the nature of the actual task, in that it requires participants to read, was implemented in the full study.

(3) *Selecting four minutes and thirty seconds as the allotment for the time pressure.* As the finding of the pilot study was that participants took on average between 4 minutes and 6 seconds to four minutes and 43 seconds, depending on the processing level, 4 minutes and 30 seconds, within this range, appears to be a more appropriate time pressure. Setting a lower time pressure would be counterproductive, as if participants realize too early that the time pressure is not a time limit, which is what occurs when time runs out and they are allowed to continue, they may no longer proceed as quickly as possible. Indeed, this appeared to be the case in the pilot, as more participants exceeded five minutes in the lower time pressure (4) than in the higher time pressure (1).

(4) *Including a supplementary time-limit for participants who fail to complete the reading within the allotted Time Pressure.* As in the pilot study, one participant in the 4 minute 30 second time pressure group took over seven minutes, a one minute thirty second time-limit countdown clock
was employed after the time pressure expired in order to ensure that participants continued to proceed as quickly as possible.

(5) *Modifying probe of participants’ processing for meaning in the Post-Test Survey.* In the pilot, participants were asked to evaluate the extent to which they were able to understand the meaning of the passage and to explain their selection. Nonetheless, participants may attempt to understand the meaning (i.e., process for meaning), but still be unsuccessful. In addition, some participants addressed the difficulty of the text in their answer to this question, indicating that this item of the survey may have been somewhat ambiguous. For these reasons, the probe of participants’ processing for meaning in the Post-Test Survey specifically addressed whether or not participants were trying to understand the meaning as opposed to whether or not they understood the text.

(6) *Stratifying.* While the pilot stratified the processing levels within working memory capacity levels, it was pointed out that since participants are randomly assigned to Processing Level and TANTA, stratifying for working memory capacity is unnecessary. In other words, it may be more important to control for Spanish level, given that Intro II and Intermediate I participants may have different levels of Spanish, and that this might confound the results. For this reason, Intro II and Intensive Basic participants were recruited towards the end of the semester, and Intermediate I and Intensive Intermediate participants were recruited towards the beginning of the semester.
(7) Increase Post-Test Survey answer choices to seven from five. Also for the last question of the Post-Test Survey, 12 out of the 15 participants selected 3 as their response. To increase the variability of the responses for this important question regarding participants processing of meaning, the number of potential choices was increased to 7.

(8) Clarifying the criteria for eliminating participants for not processing for meaning.

The pilot study used many sources to determine whether a participant was processing for meaning or not. One of those, the Practice-test score, did not appear to be related to processing for meaning in the main task, potentially because it was made up of only four questions. It was, therefore, discarded as a method of determining processing for meaning in the main study.

Total time did not provide convincing evidence that any participant was not processing for meaning. Therefore, the use of time as an indicator of processing for meaning was reconsidered in the main study, in which participants who used less than one-third the average amount of time were eliminated. In addition, time between clicks, which gives a more accurate gauge of the degree to which participants may have processed for form to the exclusion of processing for meaning, was examined.

(9) Implementing the Depth of Processing elimination criteria. Participants in depth of processing levels 2 and 3 (i.e., identifying and interpreting) were maintained in the pilot regardless of whether or not they processed the minimum number of items, with the justification that they were still processing at the required level, and that there was a limited number of participants. In the main study, however, participants that did not reach the minimum criteria of
the processing levels were eliminated in order to establish a baseline consistency of processing at the level of identifying and interpreting.

(10) *Adding an additional research question.* The main study, in addition to the six research questions explored in the pilot, also examined the impact of amount of processing (i.e., number of clicks) on comprehension. This was to explore the possibility that participants that more consistently processed for form may have been more affected, either positively or negatively, in terms of their performance.

(11) *Modifying the statistical analyses.* First, as the main study did not stratify by working memory capacity, the main study controlled for working memory capacity by initially employing an ANCOVA, instead of an ANOVA, for research questions I, II, and III. As Working Memory Capacity was not found to be significantly related to comprehension, however, an ANOVA was ultimately employed. Second, with regard to research question V, a binominal correlation was performed in order to directly assess the relationship between working memory capacity and reading comprehension.

(12) *Monitoring for returning to earlier portions of the text.* In the main study, the screen recording was used to monitor for rereading or returning to click. Rereading was defined as returning to considerably earlier portions of the text and subsequently reading substantial portions of the text again. Returning to click was defined as returning to considerably earlier portions of the text to engage in substantial processing for form.
Introduction to the Present Study

Participants

One hundred and thirteen students enrolled in Intro II, Intensive Basic, Intermediate I, and Intensive Intermediate Spanish participated in both sessions of the study. In order to control for level, students enrolled in Intro II and Intensive Basic Spanish participated in the second session of the experiment towards the end of the course, while participants in Intermediate I and Intensive Intermediate took part in the experiment towards the beginning of the course. Students in Intro II/Intensive Basic, Intermediate I, and Intensive Intermediate I Spanish classes had been exposed to approximately 70, 82, and 95 class hours, respectively, when they took part in the second session of the study (the session that examined Spanish comprehension). Participants were randomly assigned to six groups divided by three depth of processing levels and Think-Aloud vs. Non-Think-Aloud groups (TANTA). Twelve participants were first eliminated for taking more than 15 additional seconds after the countdown timer had expired. Of the remaining 101 participants, twenty-one were excluded for not processing sufficient grammatical forms, not processing at the appropriate level, clicking on too many non-target words, not processing for meaning, rereading, being under 18 years of age, or a combination of the preceding, leaving 80 total final participants. Non-eliminated participants included fifteen students with advanced knowledge of languages other than Spanish and English and native languages other than English, as well as two individuals outside the college age range of 18-22. One participant with a hearing impairment was maintained in the study, as success was not dependent on hearing ability.
Linguistic Target

The linguistic target of the pilot is the Spanish past aspect. In Spanish, as in many other romance languages (e.g., French, Portuguese), the past is expressed with either the imperfect or the preterit tense, based on the meaning and the context. With the exception of three irregular verbs (\textit{ser} “to be”, \textit{ver} “to see”, \textit{ir} “to go”), the imperfect aspect follows a predictable pattern, ending in \textit{aba}, \textit{abas}, \textit{aba}, \textit{ábamos}, \textit{abais}, \textit{aban} for \textit{-ar} for verbs whose infinitives end in \textit{-ar}, and in \textit{ía}, \textit{iás}, \textit{iá}, \textit{íamos}, \textit{íais}, \textit{ían} for verbs whose infinitive ends in \textit{-er} or \textit{-ir}. The preterit aspect is very irregular, but its form can still be clearly distinguished from the imperfect as it never ends in \textit{-ía} or \textit{-aba}. The Spanish past aspect is ideal for the current study because it neatly allows for two depths of processing: Identifying, meaning that the participant identifies that the verb is in the past; and Interpreting, meaning that the participant not only notes that the verb is in the past tense but makes a mental note of which aspect of the past tense the verb is in.

Experimental Texts

Two reading passages were used in the study, \textit{Frida Kahlo} and \textit{Los Aztecas} (‘The Aztecs’). The first of these two reading passages, \textit{Frida Kahlo}, served as a practice exercise to ensure that students understood the activity.

The Practice Reading Passage, \textit{Frida Kahlo}, is a modified version of an authentic biography on Frida Kahlo found on the cultural section of the family-oriented website Padres Hispanos (‘Hispanic Parents’). The text was shortened to 171-words long in order to reduce time spent on the practice activity. The passage was also modified to equally distribute the aspect of the past tense verbs between the imperfect and the preterit tenses. In total, the text includes 14 target forms, 7 in the preterit tense and 7 in the imperfect (Appendix A).
The Main Reading Passage, *Los Aztecas*, was first modified by Leow et al. (2008), with additional modifications being made for the current pilot. *Los Aztecas* is 373-words long and is about the ancient Aztec civilization and their much revered Sun God, Tonatiuh. As in the case of the practice passage, *Los Aztecas* was modified to equally distribute the new target forms, the imperfect and the preterit. Additional minor changes were also needed to allow for these additional exemplars of the target form to be semantically and contextually appropriate (Appendix B).

Both texts were computerized in order to control for time and backtracking. Participants were also screen-recorded. Participants in Level 2 were instructed to click on verbs in a past tense, and participants in Level 3 were required to click once for imperfect verbs and twice for preterit verbs (cf. Procedure).

**Materials**

**Think-Aloud Tasks.** The participants in the Think-Aloud group were told to think aloud, which was explained to them in an instructional video, which borrows from Goo’s (2010) Think-Aloud instructions, and also provides an example of the researcher thinking aloud (Mercer & Adrada, unpublished materials).

Participants then performed a Think-Aloud practice task explained to them in the video instructions, consisting of verbalizing their thoughts while they read a short excerpt of *Dos palabras* (Mercer & Adrada, unpublished materials; see Appendix G). Written instructions preceded both the practice and the actual reading indicating to participants that they needed to think-aloud. Participants were also given reminders, when necessary, to think aloud.
Assessment Materials

**Brief Surveys.** Participants took two brief surveys, one at the beginning and one at the end of the experiment. The Pre-Test Survey ([Appendix C](#)) asked participants how old they were, what languages they began to speak between different age intervals, and whether or not they had any non-corrected hearing or visual impairments. The Post-Test Survey ([Appendix H](#)) asked participants what percentage of the Aztecs text they had known prior to reading, and what percentage of the comprehension questions this information was helpful to answer. The Post-Test Survey also asked participants to rate their effort on the activity, to evaluate the degree to which they were trying to understand the meaning of the Aztecs passage, as well as to explain their ranking, and finally, to note what they focused on while they were reading the text.

**Operation Span Task.** Goo’s (2010) version of the OSPAN task, “originally developed by Turner and Engle (1989)” (Goo, 2010, p. 724), was used in this study, with the addition of a 4 minute 10 second instructional video developed by the current researcher. In the instructional video, participants heard an explanation of the task, which visually includes Goo’s (2010) written PowerPoint instructions, and then saw an example of the researcher taking the practice Operation SPAN task. After watching this video, the researcher asked participants if they had any questions, and subsequently allowed them to proceed to the practice OSPAN task. After taking the practice task, which included six mathematical operation-word pairs divided evenly into three sets, participants were asked if they had any remaining questions and were allowed to proceed to the actual, to-be-analyzed, OSPAN task. The actual task included forty-two mathematical operation-word pair slides, each of which displayed one mathematical operation (e.g., 2+8 =10) followed by a monosyllabic English word (e.g., BANK). The participants were instructed to read aloud each mathematical operation-word pair. Following each operation-word
pair slide, another slide was presented asking the participant to mark on their answer sheet whether the mathematical equation was correct or not. The number of operation pairs randomly varied from 2 to 5 in each set. After each set, a slide prompted the participants to recall the words presented after each mathematical operation in the set. They were “given 8 seconds for set size 2, 11 for set size 3, 14 for set size 4, and 17 for set size 5” (Goo, 2010, p. 725).

**Multiple-Choice Comprehension Tasks.** Participants took two Multiple-Choice Comprehension Tasks: The Multiple-Choice Comprehension Practice Task and The Multiple-Choice Comprehension Task.

The Multiple-Choice Comprehension Practice Task consisted of four questions based on the content of the Practice Reading Passage, *Frida Kahlo* (Appendix E).

The Multiple-choice Comprehension Task was a modified and expanded version of the Multiple-Choice Task in Leow et al. (2008), which contained multiple-choice questions based on the content of the Main Reading Passage, *Los Aztecas*. Modifications were made to ensure that the necessary changes made to the text were reflected in the questions, and that the number of multiple-choice questions per paragraph corresponded to the number of target forms. For example, a paragraph with two imperfect and two preterit forms (or four total past forms) also contained the information necessary to answer four comprehension questions. This doubled the total number of multiple-choice questions included in Leow et al. (2008) to 20 (Appendix F).

Both the Multiple-Choice Comprehension Practice Task and the Multiple-Choice Comprehension Task were administered electronically on Heroku, a cloud platform for applications, to maintain a consistent electronic format throughout the second session of the experiment.
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

Procedure

In the first session, a survey and an operation working memory span task were administered to participants. The operation working memory span task was preceded by 3 practice sets, which were not part of the final data analysis. The same operation span task that Goo (2010) used was employed in this study, with only minor editing. Prior to the second session, participants were randomly assigned to six conditions:

1. Depth of Processing Level 1: Control (Reading for meaning only); Think-Aloud group (17)
2. Level 2: (Reading for meaning and processing forms at the depth of identifying); Think-Aloud group (13)
3. Level 3: (Reading for meaning and processing forms at the depth of interpreting); Think-Aloud group (9)
4. Level 1: Control (Reading for meaning only); Non-Think-Aloud group (13)
5. Level 2: (Reading for meaning and processing forms at the depth of identifying); Non-Think-Aloud group (13)
6. Level 3: (Reading for meaning and processing forms at the depth of interpreting); Non-Think-Aloud group (15)

The participant numbers are detailed in Table 5.
In the second session, all participants were screen-recorded, while participants in the TA group were also voice-recorded. Additionally, participants in the TA group first watched an instructional video on thinking aloud, including an example of a Think-Aloud, and subsequently completed a practice Think-Aloud task (see Appendix G). Participants in the TA group were instructed to Think-Aloud throughout the experiment. Participants in both the TA group and the NTA group then listened to video instructions that explained the remaining tasks. These instructions emphasized the need to process the meaning of the text, to proceed as quickly as possible, and to read the text only once. For Levels 2 and 3, the instructions also explained to participants that they needed to process the text for form, and that they needed to do so simultaneously as they processed the text for meaning (see Appendix I). They subsequently read the modified version of the practice passage Frida Kahlo. Participants in Levels 2 and 3 completed an additional task while they read. Participants in Level 2 were required to click on past forms in the passage in order to induce processing at the level of identifying. In order to induce processing at the level of interpreting, participants in Level 3 were additionally required to label each verb in the past form as either the preterit or the imperfect by clicking once (imperfect) or twice (preterit). In order to enable clicking, computer software for the experiment was developed by Gregory S. Mercer, Senior Software Engineer at the Stanford Graduate School.
of Business and the researcher’s father. The software highlighted and recorded the words participants clicked on. When the countdown pressure timer reached 1 minute, it turned yellow to warn students that time was running out. For participants that allowed the timer to reach zero, a pop-up screen appeared asking them whether or not they finished the reading. Participants that indicated that they had not finished the reading at this point were redirected to the reading and given an additional 1 minute and 30 seconds to finish. This countdown time limit appeared in red and began to flash when it reached 20 seconds alerting participants that they were nearly out of time. For participants that allowed this extra time to run out, a pop-up screen again appeared asking them if they had finished the reading, and they were allowed to continue if they indicated that they had not. When it appeared that participants forgot that they needed to click, participants were reminded to do so, and told to start at the point they were at in their reading. When they finished reading, participants then completed the Multiple-Choice Comprehension Practice Task. Participants were not allowed to return to the text once they began the Multiple-Choice Comprehension Practice Task.

After completing the Multiple-Choice Practice Task, participants read the Aztecs passage, and subsequently took the Multiple-Choice Comprehension Test, under the same conditions that they performed the practice tasks. At the conclusion of the experiment, participants took the Post-Test Survey.

Elimination Criteria

Fifteen participants in Levels 2 and 3 were excluded from the analysis for not processing at their assigned depth of processing. Two participants were eliminated for being outliers in terms of the number of non-target forms they clicked on, namely, 8 and 15 compared to a range
of 0 to 4\textsuperscript{xvi}. Participants in Level 2 were excluded for not processing minimally at the level of identifying. Similar to previous studies (e.g., Leow et al., 2008; Morgan-Short et al., 2012; VanPatten, 1990), identifying (previously referred to as noticing/circling, as in Leow et al., 2008) was operationalized as clicking on at least 60% (12 out of 20) of the target forms. Participants in Level 3 were excluded if they did not process at the depth of interpreting. Processing at the depth of interpreting was operationalized with two criteria. First, participants needed to minimally click on (identify) 60% of the target forms. Second, in order to control for guessing, which would only involve identifying the target form, participants needed to correctly interpret at least 75% of the forms they clicked on. Seventy-five percent was chosen as the threshold because students who label 80% of the target forms (mid-way between the 60% minimum and the 100% maximum) have less than a 5% chance of correctly guessing 75% or more of the forms they labeled\textsuperscript{xvii}.

Five participants were eliminated for not processing for meaning while reading the text. For the Think-Aloud group, the concurrent verbal reports, which include a screen recording, and answers in the Post-Test Survey were analyzed to determine if participants were processing for meaning. With regard to the Non-Think-Aloud group, as concurrent verbal reports are not available, the screen recording\textsuperscript{xviii}, and answers in the Post-Test Survey were used to make this determination. Time between clicks, in the case of participants processing for meaning and form, and the Screen Recording, were considered as part of the determination as well. Time between clicks was considered a more accurate gauge of processing for meaning than time on task as it also provided data on the degree to which participants processed for form at the exclusion of processing for meaning. For all processing groups, total time was also considered, however, only in the instances in which total time was unreasonably low, operationalized as less than one-third the average time spent by participants in the same TANTA group.
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

Two participants in the processing for meaning and form at the depth of interpreting commented on the aspect of one of the forms in the text. Given that they only commented on one of the forms, however, they were not eliminated as they were considered to have primarily processed for form at the depth of identifying.

Participants that did not finish the story when the time-pressure was up were not eliminated. This is because the purpose of the time-pressure was not to ensure that all participants finish the tasks in the same or approximately the same amount of time, but, rather, to encourage participants to complete the tasks as quickly as possible. However, when participants run out of time on the time-pressure, a one minute 30 second time-limit countdown clock began in order to see to it that participants continued to proceed as quickly as possible. Twelve participants who did not finish within fifteen seconds after this additionally allotted time was up were eliminated, as it was viewed as likely that these participants were not following the instructions to proceed as quickly as possible.

Two questions in the Post-Test Survey pertained to participants’ prior knowledge. The first asked participants what percentage of the text they had known prior to reading, and the second asked participants what percentage of the text their prior knowledge helped them to answer. Given that prior knowledge pertinent to the comprehension questions has the potential to influence the results, further examination was given to the five participants who noted that their prior knowledge was greater than 30% for potential elimination. One participant answered 100% to the question, however, his concurrent verbal report revealed that he had misunderstood the question to mean what percentage of his knowledge after reading the text was helpful to answering the questions. Another participant, who answered 40%, noted that he “knew about how they sacrificed people and how they were religious.” This comment, however, seems to
reflect general common knowledge about the Aztecs that was unlikely to be helpful to answering the questions. Additionally, two participants answered 50%, but only received a 55% score, which seems to indicate that they exaggerated the extent of their prior knowledge, as it is unlikely that they only comprehended 5% of the text. While the remaining participant, who answered 40%, did not give an indication in his concurrent verbal report as to whether he actually had this amount of prior knowledge, given the unreliability of the answers of the four participants just mentioned, it was determined that the question was unreliable, and no participant was eliminated for prior knowledge. Evidence in the concurrent verbal reports revealed that some participants who answered less than 30% had prior knowledge to answer two to three questions, however, as this only represents 15% of the questions, it was determined to not be sufficient to warrant elimination.

One participant engaged in rereading, defined as returning to considerably earlier portions of the text to read a substantial amount of the text again. However, this participant had already been eliminated for not processing for form. No participant was found to have engaged in returning to click, defined as returning to considerably earlier portions of the text to substantially process for form. One participant that came close, the same participant that was found to have reread, clicked on two forms in considerably earlier portions of the text.

**Scoring**

**Multiple-Choice Comprehension Tasks.** The Multiple-Choice Practice Comprehension included four questions while the Multiple-Choice Comprehension Task included twenty. As some participants skipped portions of the text and/or did not complete the text, participants were
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

Given a percentage score based on the total number of questions that pertained to the sections of the text that they had read (see Skipping).

**Working Memory Capacity.** The Working Memory Span Scores was calculated by combining the processing and recall portions of the Working Memory Span Task. The processing part of the working memory tasks was scored by giving one point to each correct answer and computing a percentage for each working memory span task. The recall portion of the working memory tasks was scored following the partial-credit load scoring procedure, as in Conway et al (2005) and Goo (2010), dividing the number of recalled items by the total number of items (following Goo, 2010; Conway et al., 2005). Working Memory Score was then calculated by averaging the percentage recall score with the percentage processing score on the OSPAN.\textsuperscript{xix}
CHAPTER 4: RESULTS

Qualitative Analysis

In this qualitative analysis, I begin with a discussion of the elimination of participants for not processing for meaning. First, I discuss the concurrent verbal reports and the Post-Test Survey. These two were considered to be the most reliable indicators of participants having processed for meaning. Second, I discuss the use of Time Data, the screen recordings, and Multiple-Choice Comprehension scores as supplementary sources to make this determination. These were considered especially in the cases in which the Post-Test Survey did not provide a clear indication of whether or not participants had processed for meaning. In addition, I explore instances in which participants were determined to have skipped portions of the text primarily through the use of the Think-Aloud data and the screen recordings. Finally, I discuss the strategies and comments of participants both in the concurrent verbal reports and in the Post-Test Survey.

Processing For Meaning

Concurrent Verbal Reports. The concurrent verbal reports were effective in determining that participants had processed for meaning. In every Think-Aloud protocol, evidence was found to suggest that participants had processed for meaning. In all Think-Aloud protocols, participants translated portions of the text into English, which was taken as evidence that they were processing for meaning, regardless of whether or not the translation was accurate. As can be seen in the following examples, the exact translation strategy was wide-ranging, from reading the text
in Spanish and subsequently translating it (Participant 1), to translating parts of the text and reading others (P2), to directly attempting to translate the text into English (P3).

Participant 1 Los Aztecas, the Aztecs, el pueblo Azteca, the Aztec town, como pueblo primitivo primitive town descubrió una solución por los problemas presentados por las fuerzas de la naturaleza. Um found a solution to the problems presented to the natural world, natural forces.

P2 Los Aztecas. The house of the aztecs however primitive the house was primitive descubrió una solution to the problems presented por las fuerzas of nature.

P3 The Aztecs constructed a lot of monuments like they constructed a lot of monuments in order to honor the sun mm no in order to honor the sun.

Some participants also made comments pertaining to the meaning of the text, which was also taken as evidence of processing for meaning. Here is one example:

P4 This paragraph talks about how the ancient Mexicans um language of the sun the sun was important to their culture sorry to feed the sun

Post-Test Survey. The Post-Test Survey was used to determine whether or not participants were processing for meaning. Due to the Concurrent Verbal Reports providing direct evidence that participants processed for meaning, the Post-Test Survey data for Think-Aloud participants, while reviewed, was ultimately not needed to make the determination. The Post-Test Survey asked participants to rate the degree to which they tried to understand the meaning of the text on a scale of 1 to 7, as well as to explain their rating. It also asked participants what they focused on during the reading. Ratings ranged from 3 to 7. Ratings of a 5 or higher were considered
evidence of reading for meaning. While all explanations were reviewed, particular attention was given to those that had rated their attempt as a 3 or a 4. Four participants were eliminated for not processing for meaning based on their answers to the Post-Test Survey, two of which follow:

P5 I got distracted during the middle and failed to focus on the meaning of the content (emphasis added). I felt that I understood the beginning and the end better.

P6 I was concerned about the timer and rushed through it, stupidly. I also got caught up on finding the verbs and was then unable to go back and understand the meaning.

In the first of these two cases, the participant directly indicates that he did not process for meaning, while in the second it is understood that the participant did not do so as he laments not being able to go back to “understand the meaning.” The following are examples of Post-Test Survey responses from participants that suggested that they had processed for meaning.

P7 It was difficult, but something I am interested in and so I was trying to comprehend and understand, but there was vocab that I didn’t know.

P8 While trying to comprehend and look out for the past tense words it was hard to fully comprehend, especially when looking at the time and how much you have left.

P9 I tried to read and understand just as hard as I do in Spanish class--though I don t know every word, I try to figure out what everything means through context.

P10 I was trying to understand the English translation, and was hung up on translating rather than understanding.

Participants 7, 8, and 9, while noting that the task was difficult, nonetheless indicates that he made an effort to understand the meaning. While P10 indicates that he was focused on translation, this was interpreted by the researcher as being a form of trying to understand the text.
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

Some participants’ post-test survey responses did not give a clear indication as to whether or not they were processing for meaning, the following is one example.

P11 I tried to understand but I caught myself focusing on the verbs more and picking up minimal details.

This participant both indicates that he tried to understand, while at the same time giving the indication that he may have focused just on the verbs. For participants such as this one whose Post-Test Survey did not give a clear indication of whether or not the participant had processed for meaning, Time Data, the screen recordings, and Multiple-Choice Comprehension scores were also taken into account to make this determination.

**Time Data.** In those cases in which the Post-Test Survey provided inconclusive data, time between clicks was used to further explore whether or not participants were focusing on forms to the exclusion of processing for meaning. Nine seconds, or just over two-thirds the average amount of time after processing for form eliminations and total time eliminations, was considered to be sufficient time to provide an indication that the participant was processing for meaning. This was corroborated by the fact that a participant eliminated by the Post-Test Survey had spent less than 9 seconds between clicks, although it should be noted that 2 participants who were coded for processing for meaning based on the Post-Test Survey also took less than 9 seconds. Returning to the previous example (P11), this participant took approximately 15 seconds between verbs, and, for this reason, was judged to have been processing for meaning, as merely searching for forms would not require this much time between forms. Total time on task was used to eliminate participants that had taken too little time to have been able to process for meaning. Less than one-third the average amount of time spent on the task was determined to be
the threshold for elimination, for which one participant was eliminated. Total time on task was
also used to eliminate participants that took too much time, and therefore were potentially not
proceeding as quickly as possible. More than 15 seconds beyond after the end of the second
timer was determined to be the cutoff point, eliminating 12 participants.

**Screen recordings.** In conjunction with the time data, the screen recordings were used
particularly in those cases in which the Post-Test Survey did not give a clear indication of
whether or not the participant had processed for meaning. Nonetheless, following the text with
the cursor did not provide strong evidence that the participants were processing for meaning. In
one instance in which the participant moved the cursor quickly across the text, the multiple-
choice comprehension score was reviewed to determine if the participant was processing for
meaning. It was found that this participant had a slightly above average score, and for this reason
was coded as having processed for meaning.

**Multiple-Choice Comprehension.** Due to the potential of eliminating participants that
attempted to read the text for meaning but were ultimately unsuccessful, low scores were not
considered. However, scores higher than fifty-three percent, the average score prior to
eliminations (it was also the average after eliminations), were considered evidence of processing
for meaning.

**Skipping**

Concurrent verbal reports, in conjunction with the screen recordings, were also used to determine
if participants had read the entire reading, and, if not, which portions of the text they had
skipped. Secondarily, participants’ answers to the questions pertaining to these sections were reviewed to make this determination. Questions related to sections that were skipped were omitted for each participant. Among non-eliminated participants, a total of twelve participants had skipped portions of the passage, with many more participants having skipped among those in the Processing for Meaning Only Condition (7), than those in the Processing for Meaning and Form at the Depth of Identifying (2), or in the Processing for Meaning and Form at the Depth of Interpreting (3), with two participants in the processing for meaning only group skipping the entire last paragraph because they did not scroll down to notice it and the timer ran out just when they finished the second-to-last paragraph. However, when those participants that had been eliminated for not having processed sufficiently (at least 60%) or deeply enough (for the Processing for Meaning and Form at the Depth of Interpreting Condition) for form, an elimination that did not occur in the Processing for Meaning Only Condition, the difference in amount of skipping across Conditions is no longer substantial (7 vs. 4 vs. 6). As can be deduced here, the percentage of participants who skipped sections of the reading among that did not meet the minimum processing for form requirements was much higher than that among non-eliminated participants (33% vs. 15%). Among non-eliminated participants, skipping was rare in the Non-Think-Aloud group, with only one individual having skipped among non-eliminated participants, compared to eleven in the Think-Aloud group.

Time Eliminations

Participants eliminated for taking more than 15 seconds beyond the one minute and thirty second countdown timer were not equally distributed across the six groups. Only two participants were eliminated for this violation in the Non-Think-Aloud group, with the remaining ten belonging to
the Think-Aloud group. In addition, only one participant exceeded the amount of time allotted by more than 15 seconds in the Processing for Meaning Only group, while four did so in the Processing for Meaning and Form at the Depth of Identifying, and seven did so in the Processing for Meaning and Form at the Depth of Interpreting Condition.

Participant Strategies and Commentary

Think-Aloud Data. As noted in the discussion of the use of the concurrent verbal reports to determine if participants are processing for meaning, all participants translated portions of the text in order to gain an understanding of the text. The frequency of translation, however, varied between participants. In addition to this strategy, many participants also made comments in their concurrent verbal reports. Some participants, for example, commented on their difficulty with the vocabulary of the text. Here are some instances in which that occurred:

P12 corazón, I should know what that is

P13 In with the Aztecs there was some primitive I don’t know what that is but it’s preterite (referring to descubrió)

P7 madurez don’t know what that means

Participants occasionally attempted to guess at the meaning of the word they did not know, which normally occurred in the form of translation, but in some instances as comments, as in the following examples.

P14 defectos de los humanos they mean defectors?

P15 señor de los I’m guessing that’s what rain is? cielos

Several participants commented on the nature of the forms they were asked to process, as in the following instances:
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

P16 daba that’s imperfect

P17 incluian that’s preterite oh imperfect

In addition to translating and/or commenting, most participants also read portions of the text in Spanish. Unlike in the pilot, in no instance did participants in the Think-Aloud group limit themselves to only reading the text aloud in Spanish.

Post-Test Survey. Many participants commented on the difficulty of processing for form and meaning simultaneously, and on the challenge posed by the countdown timer. The following are some examples.

P18 It was a balancing act between reading quickly and understanding the essay.

P19 I struggled to focus on the reading because I was trying to look for the correct verb forms as well as focus on the reading. It was hard to do both in the specific time period. Also, seeing the clock made me nervous and want to read faster which led to lower concentration.

P20 Due to the time pressure, I mainly tried to focus on finding the past tense verbs, and skimmed over the information. As I got closer to the end, I realized how much information there was, and knew I would not be able to remember it all in time.

Participant 18 commented that time-limit had an impact on their strategy, in that it forced them to read more quickly, suggesting that the use of the time-limit may have been effective in ensuring that participants proceeded as quickly as possible. Participant 19 noted the struggle of processing for form and meaning “in the same time period,” and that the time-limit had an impact on his “concentration.” What this participant may have been referring to is that the
limited amount of time available restricted the cognitive resources that he had available to him. Participant 20, who was eliminated, appears to indicate that he was sufficiently impacted by the need to process for meaning and form simultaneously that he was unable to complete the task successfully.

Quantitative Analysis

Preliminary Analysis

While Intro II/Intensive Basic and Intermediate I/Intensive Intermediate Spanish learners participated in the experiment towards the end and the beginning of their course, respectively, in order to reduce the difference in level between these two groups as much as possible, a t-test was nonetheless performed to ensure that the slight difference in exposure to Spanish would not have an impact on the results. No significant difference was found between Intro II/Intensive Basic and Intermediate I/Intensive Intermediate Spanish learners, $t(78) = .151, p = .880$. Values for the analysis can be seen in Table 6.

<table>
<thead>
<tr>
<th>Spanish Level</th>
<th>Score (n = 80)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intro II/ Intensive Basic</td>
<td>53.52 (10.72)</td>
</tr>
<tr>
<td>Intermediate I/ Intensive Intermediate</td>
<td>53.10 (13.71)</td>
</tr>
<tr>
<td>Total</td>
<td>53.37 (11.80)</td>
</tr>
</tbody>
</table>

To ensure that participants processing for form and meaning simultaneously did not take additional time to compensate for the added burden of processing for form, a 2x3 ANOVA was
performed with Think-Aloud vs. Non-Think-Aloud (TANTA) and Depth of Processing condition (DP) as potential factors, and Time in Seconds as the dependent variable. While TANTA was found to be significant, $F(1, 74) = 79.602, p = .000$, DP, $F(2, 74) = 2.154, p = .123$, was not.

**Research Questions I, II, III**

In order to examine the first three research questions which explore the impact of Depth of Processing for Form (control vs. identifying vs. interpreting) and TANTA (Think-Aloud vs. Non-Think-Aloud), and the interaction between Depth of Processing and TANTA on comprehension, a 2-way ANCOVA was performed with Depth of Processing (DP) and TANTA as independent factors, Working Memory Capacity (WMC) as the covariate, and Multiple-Choice Comprehension (MCC) as the dependent factor. A main effect was found for TANTA, $F(1, 73) = 13.465, p = .000$, and for DP, $F(2, 73) = 3.207, p = .046$, but not for Interaction, $F(2, 73) = 0.962, p = .387$. As the covariate WMC was also not found to be significant, $F(1, 73) = .001, p = .972$, an ANOVA with the same independent and dependent factors was performed to further examine the results. As can be seen in Table 7, DP and TANTA were again found to have a main effect, while interaction was not. Table 8 and Figure 1 detail descriptive statistics for the ANOVA.

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>$F$</th>
<th>$\eta^2$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP</td>
<td>2</td>
<td>3.412</td>
<td>.071</td>
<td>.038</td>
</tr>
<tr>
<td>TANTA</td>
<td>1</td>
<td>14.004</td>
<td>.145</td>
<td>.000</td>
</tr>
<tr>
<td>DP*TANTA</td>
<td>2</td>
<td>.974</td>
<td>.020</td>
<td>.382</td>
</tr>
</tbody>
</table>

Table 7
*Impact of Depth of Processing and TANTA on Comprehension*
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

Table 8

<table>
<thead>
<tr>
<th>DP Level</th>
<th>TANTA (n = 39)</th>
<th>NTA (n = 41)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Control)</td>
<td>60.11 (10.26)</td>
<td>55.09 (11.77)</td>
</tr>
<tr>
<td>2 (Identifying)</td>
<td>57.94 (10.49)</td>
<td>45.00 (10.00)</td>
</tr>
<tr>
<td>3 (Interpreting)</td>
<td>55.35 (12.74)</td>
<td>46.33 (8.76)</td>
</tr>
<tr>
<td>Total</td>
<td>58.29 (10.81)</td>
<td>48.69 (10.87)</td>
</tr>
</tbody>
</table>

Figure 1

Descriptive Statistics
The mean values in Table 8 show that thinking aloud has a positive impact on comprehension. A post-hoc Scheffé analysis was performed for Depth of Processing, and the significant difference found was between Processing for Meaning and Form at the Depth of Interpreting and Processing for Meaning Only (p = .022). There was no significant difference in performance between Processing for Meaning and Form at the Depth of Identifying and Processing for Meaning Only (p = .080). Table 9 presents the descriptive statistics for the Scheffé Analysis.

Table 9  
Post-Hoc Scheffé Analysis

<table>
<thead>
<tr>
<th>Control</th>
<th>Identifying</th>
<th>Interpreting</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.93&lt;sup&gt;a&lt;/sup&gt;</td>
<td>51.47&lt;sup&gt;ab&lt;/sup&gt;</td>
<td>49.72&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>(11.03)</td>
<td>(12.01)</td>
<td>(11.09)</td>
</tr>
</tbody>
</table>

In summary, Depth of Processing and TANTA were found to have a significant impact on comprehension, while Interaction between Depth of Processing and TANTA did not. With regard to Depth of Processing, the Post-Hoc Scheffé Analysis revealed that the only significant difference was between the two extreme depth of processing groups, namely, depth of processing for form at the level of interpreting while simultaneously processing for meaning and processing for meaning only. For TANTA, thinking aloud had a positive impact on comprehension.

Research Question IV

Research question IV, namely, *Is there a relationship between amount (number of clicks) of processing for form and L2 comprehension score?*, was explored by running a binomial correlation between number of clicks and Multiple-Choice Comprehension score (MCC) for
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

participants processing for form at the depth of identifying or interpreting. The relationship
between number of clicks and MCC was not found to be significant, $r = .250$, $p = .08$.

Research Questions V, VI, VII

Research question V examined whether or not there was a relationship between Working
Memory Capacity (WMC) and Multiple-Choice Comprehension (MCC). A binomial correlation
was performed to examine this question. In the analysis, no relationship was found between
WMC and MCC, $r = .010$, $p = .933$. Research questions VI and VII examined whether there was
an interaction between Working Memory Capacity and Think-Aloud vs. Non-Think-Aloud
(TANTA) in their relationship to Multiple-Choice Comprehension, or an interaction between
WMC and Depth of Processing Level (DP) in their relationship to Multiple-Choice
Comprehension. A multiple regression analysis, with DP, TANTA, WMC, WMC*DP, and
WMC*TANTA as independent variables, and Multiple-Choice Comprehension as the dependent
variable, was performed to investigate these questions. While DP and TANTA were found to be
significant predictors, WMC, WMC*DP, and WMC*TANTA were not. The regression equation
found was $RMCC = 56.674 + 8.543TA - 3.699DP + .386WMC - .044WMC*DP - .558WMC*TA$,
with $R$ squared equal to .250. The values for this analysis can be found in Table 10. The
significant relationship between DP and MCC, as well as that between TANTA and MCC, can
be seen in Figure 2.
Table 10
Coefficient Variables of Multiple Regression Analysis

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP</td>
<td>-3.699</td>
<td>-2.460</td>
<td>.016</td>
</tr>
<tr>
<td>TANTA</td>
<td>8.543</td>
<td>3.493</td>
<td>.001</td>
</tr>
<tr>
<td>WMC</td>
<td>.386</td>
<td>3.493</td>
<td>.016</td>
</tr>
<tr>
<td>DP * WMC</td>
<td>-.004</td>
<td>-.206</td>
<td>.837</td>
</tr>
<tr>
<td>TANTA * WMC</td>
<td>-.558</td>
<td>-1.634</td>
<td>.106</td>
</tr>
</tbody>
</table>

Figure 2

In Figure 2, the downward slope of the blue line and the green line, respectively representing the Non-Think-Aloud group and the Think-Aloud Group, reveals a negative correlation between Depth of Processing Level and Multiple-Choice Comprehension. In addition, the gap between the two lines shows that the thinking-aloud is correlated with higher Multiple-Choice Comprehension scores. As the gap appears to widen at greater depths of processing for form, a
bivariate correlation was performed between TANTA*DP and MCC. Interaction was not found to be significant, $r = .045$, $p = .689$.

A post-hoc t-test was performed to explore the possibility that effort might have played a role in the significant positive impact of thinking aloud. The t-test revealed that participants in the TA group did indeed rate their effort as significantly higher, $t(78) = -3.014$, $p = .003$. 
CHAPTER 5: DISCUSSION AND CONCLUSION

Introduction

The principle goal of the current dissertation was to assess the impact of depth of processing for form while simultaneously processing for meaning on reading comprehension. The dissertation also sought to examine the effect of thinking aloud on comprehension, and, more specifically, on the potential interaction between thinking aloud and depth of processing for form. Additionally, the possible relationship between amount of processing for form and comprehension was analyzed. Finally, given the potential for individual processing abilities to have an impact, the role of working memory capacity was also explored. The current chapter discusses the results of the dissertation in terms of each of the seven research questions, suggests potential pedagogical implications, notes limitations to the study, recommends future research, and outlines the potential conclusions that can be drawn.

Discussion

Research Question I

The answer to the first research question, *Does Depth of Processing for Form (control vs. identifying vs. interpreting) while processing an L2 text for meaning have an impact on comprehension?*, can be answered partially in the affirmative in the sense that only the highest level of depth of processing for form negatively impacted performance. The results of the ANOVA showed a significant main effect for Depth of Processing with a medium effect size ($n^2 = .071$; following Cohen, 1988). Additionally, answers to questions in the Post-Test Survey
revealed that many participants specifically cited processing for form as distracting, or one that made processing for form more difficult. The Post-Scheffé analysis in Table 9 revealed a significant difference between processing for meaning and form at the depth of interpreting and processing for meaning alone, but no such significant difference between processing for meaning and form at the depth of identifying and processing for meaning only. This appears to indicate that a minimum depth of processing for form is required to disrupt comprehension, supporting Leow, Hsieh, and Moreno’s (2008) original inkling that greater depths of processing might be necessary in order to have the “hypothesized detrimental effect” of VanPatten’s (2004) Primacy of Meaning Principle (p. 688). That is to say, participants may be able to superficially process for form while they process for meaning. The level of cognitive resources required to process for form at a greater depth, however, may not leave sufficient resources for processing for meaning to be conducted at par. These findings at least partially contradict most of the previous research on Processing for Form and Meaning (Greenslade, Bowden, & Sanz, 1999; Hulstijn, 1989; Morgan-Short, Heil, Botero-Moriarty, & Ebert, 2012; VanPatten, 1990; Wong, 2001), with the arguable exception of Leow et al. (2008).

As noted in the review of the literature, Hulstijn (1989) conducted two experiments examining the impact of focusing on form and meaning on comprehension and production, as compared to groups focusing on meaning or form. In his experiment, focus on form and focus on meaning both required participants to perform a specific task. Focus on form and meaning, on the other hand, was operationalized by merely asking participants to process for meaning and form. In one of the two experiments performed, Hulstijn (1989), contrary to the current study’s finding, the processing for meaning and form condition performed significantly better than the meaning condition on comprehension. Two key differences between the two studies, however,
may explain the divergence in results. First, participants in the form and meaning condition in Hulstijn (1989) were not given a time limit, and therefore, may have taken more time to compensate for the additional task of processing for form. Second, participants in Hulstijn’s (1989) form and meaning condition did not perform a task to ensure that they were indeed processing for form, and therefore, the degree to which they processed for form may have been less than in the current study.

Disregarding modality, VanPatten (1990), Greenslade et al. (1999), Wong (2001), and the current study, with the exception of the measures employed to ensure participants processed for meaning, conducted similar comparisons between processing for meaning only and processing for meaning and form at the depth of identifying. In each of these studies, participants read or listened to a text in a limited amount of time and subsequently took a comprehension test. In addition, depending on their condition, participants either processed only for meaning or both processed for meaning and identified, by making a check mark upon hearing, circling, underlining, or, in the case of the current study, clicking on, a particular grammatical form.

While Greenslade et al. (1999), VanPatten (1990), and Wong (2001) found, in at least one of their experiments for at least one grammatical form, that participants processing for meaning only performed significantly better than those participants processing for meaning and form at the depth of identifying, the current experiment did not find evidence of a significant difference between these two conditions. Two key and related differences, however, should be pointed out between VanPatten (1990), Greenslade et al. (1999), and Wong (2001), and the current study. First, in the current study, reading for meaning was repeatedly stressed in video (start of second session or after Think-Aloud Practice Task, depending on TANTA group), and written
instructions (before each reading) as being essential, and participants were specifically instructed not to process only for the forms or only for meaning, but to process for both simultaneously. In comparison to the repeated and explanatory instructions in the current experiment, previous experiments (Greenslade et al., 1999; Leow et al., 2008; Morgan-Short, Heil, Botero-Moriarty, & Ebert, 2012; VanPatten, 1990; Wong, 2001), including VanPatten (1990), Greenslade et al. (1999), and Wong (2001), merely “instructed” participants to process for meaning. Furthermore, the current experiment, as originally done in Leow et al. (2008), eliminated participants that were not processing for meaning, unlike in VanPatten (1990), Greenslade et al. (1999), and Wong (2001). It may be then, that, in comparison to the current study, that VanPatten (1990), Greenslade et al. (1999), and Wong (2001) did not take sufficient steps to ensure that participants were processing for meaning, which potentially could, in turn, have confounded the impact of processing for form with the impact of not processing for meaning, explaining the difference in results obtained in these studies (Greenslade, 1999; VanPatten, 1990; Wong, 2001) when compared to the current study.

One might consider the target form as a potential reason for the difference in results between the current study and VanPatten (1990), Greenslade et al. (1999), and Wong (2001), given that target forms have been found to play a role in previous research (e.g., Shook, 1994; Greenslade et al., 1999; VanPatten, 1990). VanPatten (1990) and Greenslade et al. (1999) examined the impact of processing for a lexical item (inflación), a definite article (la), and the verb morpheme (-n), while Wong (2001) did so with the English equivalents for all but the verb morpheme. VanPatten (1990) and Greenslade et al. (1999) both found that only processing for the definite article and the verb morpheme had a negative impact on comprehension, while Wong (2001) only found a significant result in the aural modality for the processing for the
definite article condition. It would appear then, at least with regard to these three studies, that
target forms that are closer to the meaning of the text (i.e., the lexical items), are less likely to
negatively impact comprehension, supporting VanPatten’s (2004) postulation that participants
tend to process for “lexical items as opposed to grammatical forms when both encode the same
semantic information” (p. 9). The current study did not use a lexical item, however, and only
employed the verb morphemes of the past tense. Given that this form is most similar to the verb
morpheme (-n) found to have a significant impact on comprehension in VanPatten (1990) and
Greenslade et al. (1999), and is similar in terms of the low likelihood that it will provide meaning
to participants that could not be gained by lexical items and context, it seems unlikely that the
difference in results found between the three studies (Greenslade et al., 1999; VanPatten, 1990;
Wong, 2001) and the current one could be explained by the distinct target forms used.

It may also be that the distinct method used to measure comprehension played a role in
the results. In the current study, as in Leow et al. (2008) and Morgan-Short et al. (2012), a
multiple-choice format was employed to measure comprehension, whereas in VanPatten (1990),
Greenslade et al. (1999), and Wong (2001) participants wrote down everything that they could
remember. As Leow et al. (2008) pointed out, given the lack of multiple-choice options, a recall
test may add an additional layer of difficulty in that success is dependent not only on
participants’ comprehension of the text, but also on their memory (p. 672). It might be that
processing for form at the depth of identifying is detrimental to the strength of the connection
made when comprehending, and therefore, impacts the degree to which the text can be
remembered, but not the degree to which the text can be comprehended. Indeed, a significant
relationship between working memory capacity and depth of processing was not found in the
current study. As Leow et al. (2008) argued in relationship to their study, given that the objective
of the current strand of research is to determine the degree to which processing for meaning and form impacts comprehension, not the degree to which it impacts comprehension and memory, the multiple-choice test employed here, and therefore, the results of the current experiment, are arguably more reliable. Alternatively, it may be that there is a difference in terms of comprehension between Processing for Meaning Only and Processing for Meaning and Form at the Depth of Identifying, but that the difference is smaller, and, therefore, more difficult to detect (i.e., requires more participants) than the difference between Processing for Meaning Only and Processing for Meaning and Form at the Depth of Interpreting. In spite of the difference in results, however, VanPatten (1990), Greenslade et al. (1999), and Wong (2001), like the current study, although requiring a lower depth of processing for form, support the suggestion of VanPatten’s (2004) Primacy of Meaning Principle that processing for form can have an impact on comprehension.

Morgan-Short et al. (2012) and Leow et al. (2008) did not find evidence that processing for form had a negative impact on comprehension, and in Morgan-Short et al. (2012), a positive relationship was found between depth of processing for form and comprehension. The results of Leow et al. (2008) are similar to those of the current study, as both did not find a significant difference between participants who only processed for meaning and those that processed for meaning while simultaneously identifying (i.e., making a check mark when hearing, circling/underlining or clicking) grammatical forms. The similarity in results between this study and Leow et al. (2008) when compared to VanPatten (1990), Greenslade et al. (1999), and Wong (2001) might be attributed to the presence in both studies of a control for processing for meaning. Leow et al.’s (2008) post-hoc analysis, as it did not include the processing for meaning only group, can be compared to the analysis performed between the two processing for form
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

conditions in the current study, namely, processing for meaning and form at the depth of identifying and processing for meaning and form at the depth of interpreting. The researchers did not find there to be any relationship between depth of processing for form and comprehension, a finding that does not contradict the results of the current study, given that no significant difference between the two processing for form groups was found. Nonetheless, the results of Morgan-Short et al. (2012), who replicated Leow et al. (2008) with a large group of participants and a Non-Think-Aloud group, differed from the current experiment with respect to the impact of depth of processing. Contrasting with the lack of a significant difference between processing for form at the depth of identifying and processing for form at the depth of interpreting in the current study, Morgan-Short et al. (2012), who like Leow et al. excluded the control group from the comparison, found a positive correlation between depth of processing for form and comprehension.

Methodological differences between Morgan-Short et al. (2012) and the current study, previously articulated in the dissertation, may explain the divergence in results. First, in Morgan-Short et al. (2012) participants were neither instructed to proceed as quickly as possible nor given a time-limit. Therefore, participants in the greater depth of processing conditions may have taken additional time to compensate for the burden of processing for form. The lack of a control for time may have confounded the results. Indeed, Morgan-Short et al. (2012) notes that attentional constraints, such as the timing of input, may be a potential reason for the difference in results between written and aural studies on processing for form and meaning simultaneously. Following VanPatten’s Primacy of Meaning Principle, if participants can take longer to process for form at greater depths, it is not contrary to his principle for their comprehension to be greater, as in Morgan-Short et al. (2012). Leow et al. (2008), the study Morgan-Short et al. (2012)
replicated, found time to be near significant in a one-way ANOVA, albeit in a comparison between processing for meaning and doing so while simultaneously processing for four different target forms. Morgan-Short et al. (2012), unfortunately, did not provide time data. Morgan-Short et al.’s (2012) lack of a control for time contrasts with the current study, which employed video instructions informing participants that they need to proceed as quickly as possible, as well as a time-pressure of four minutes and thirty seconds followed by a time limit of one minute and thirty seconds. Moreover, unlike in Morgan-Short et al. (2012), the current study also controlled for time by running a 3x2 ANOVA for time, and found no significant main effect for Depth of Processing Condition.

A second crucial and likely explanatory methodological difference between the current study and Morgan-Short et al. (2012) was the method used to establish depths of processing. In Morgan-Short et al. (2012), as in Leow et al.’s (2008) post-hoc analysis, participants were coded for depth of processing level based on their Think-Aloud protocols, meaning that depth of processing was determined by the performance of the participants, as opposed to being randomly assigned by the researcher. The methodological concern with this particular approach, however, is that it allows for mediating variables to play a role. While working memory capacity was not found to be significantly related to comprehension in the current study, this does not rule out the possibility that other variables may have been involved. In other words, participants’ individual differences, such as language aptitude, as opposed to a causation relationship between depth of processing for form and comprehension, may make them more likely to both process at a higher level and comprehend more. It seems plausible, for example, that those participants that are most familiar with the meaning of the direct object lo and verb morpheme –n might be participants that are particularly skilled at languages in general, and thus, at comprehension in a foreign
language. This provides a potential explanation for Morgan-Short et al.’s (2012) finding of a significant positive relationship between Depth of Processing for form and comprehension, particularly when viewed in contrast to the results of the current experiment, which randomly assigned participants to processing levels and found no significant difference between participants that processed for form at the depth of identifying and those that did so at the depth of interpreting.

An analysis of Morgan-Short et al.’s (2012) coding scheme also suggests that another variable is at play. Morgan-Short et al. (2012) coded participants as having processed at level III, similar to processing at the depth of interpreting in the current study, even if they had only processed one of the forms at this level. As noted in the literature review, while Morgan-Short et al. (2012) does not provide data with regard to the percentage of forms that participants coded as having processed at the level of interpreting actually processed at that level, given that Morgan-Short et al. (2012) closely replicated Leow et al. (2008), it seems reasonable to refer to the latter’s findings in this regard. Indeed, given that the main difference between Leow et al. (2008) and Morgan-Short et al. (2012) was that Morgan-Short et al. (2012) included a Non-Think-Aloud group, which was not included in the analysis of depth of processing, such a comparison seems quite fair. Leow et al. (2008), as noted in the literature review, found that most participants that had processed forms at the depth of interpreting had done so for “very few items” (p. 681). It would seem very unlikely that processing a limited number of forms at a greater depth would have a significant positive impact on comprehension, particularly considering that the comprehension questions did not require comprehension of the forms. Therefore, the fact that Morgan-Short et al. nevertheless found a significant positive relationship between processing for form and comprehension suggests that another factor may have played a role.
One might point out that the current experiment’s exploration of depth of processing for form was slightly less natural than Morgan-Short et al. (2012). While both experiments, as well as previous research in the strand (Greenslade et al., 1999; Leow et al., 2008; Wong, 2001; VanPatten, 1990), operationalized processing for form by requiring participants to click on/circle/underline or check when they heard a minimum of 60% of the forms, the current experiment additionally required participants in the processing for form at the depth of interpreting condition to click on the forms either once or twice depending on whether the form was in the imperfect or the preterit, whereas Morgan-Short et al. (2012) relied on the concurrent verbal reports to code participants as processing at higher levels. It might be possible that the difference in terms of the naturalness of the experimental conditions could have had an impact on the results. Leung and Williams (2014) argue, however, referring to their study, that the artificiality of the experiment “does not compromise the relevance of the work” (p. 19). Indeed, similar to Leung and Williams’ (2014) study, the requirements of the current experiment, namely, both randomly assigning participants to processing conditions to control for mediating variables and ensuring that participants in the highest depth of processing condition processed a minimum number of forms at the depth of their condition, necessitated the slight increase in artificiality as compared to previous studies (Greenslade et al., 1999; Leow et al., 2008; Morgan-Short et al., 2012; Wong, 2001; VanPatten, 1990).

An alternative explanation for the difference in results between the current study and Morgan-Short et al. (2012) that one might point out is that target form may have played a role, as it has in previous experiments (e.g., Shook, 1994; Greenslade et al., 1999; VanPatten, 1990). Morgan-Short et al. (2012) examined the impact of processing for a lexical item (sol), a definite article (la), the direct object pronoun (lo), and the verb morpheme (-n). Among the non-lexical
forms, the direct article was arguably the least useful in terms of the meaning information that it provided beyond that which was easily ascertained through lexical items or the context of the Aztecs text. This is because, unlike in the case of the direct object pronoun and the verb morpheme, the definite article always accompanies its referent, and the text is almost entirely understandable without it. Following VanPatten’s (2004) postulation that lexical items are more likely to be processed for meaning, and his Nonredundancy Principle that “learners are more likely to process nonredundant meaningful grammatical form[s]” (p. 11), it would be expected, if target forms are indeed playing a role, that those participants that processed for the definite article at the depth of interpreting would have the lowest scores among participants that processed for a non-lexical form at this depth. Quite the contrary, however, was found in Morgan-Short et al. (2012), where these participants actually scored the highest in terms of comprehension. Moreover, given VanPatten’s (1990) postulation that lexical items are processed first for meaning, one would expect that those participants that processed for the lexical form *sol* at the highest depth of processing would have the highest scores in terms of comprehension among those that processed for a form at this depth. In Morgan-Short et al. (2012), however, their comprehension scores were the second lowest in this group. In addition, in Morgan-Short et al. (2012), participant scores in Level 1 (similar to processing for form at the depth of identifying in the current study), and Level 2, with one exception, were lower than participant scores in Level 3 (comparable to processing for form at the depth of interpreting) for all forms. Therefore, while target form may play a role in processing for form and meaning simultaneously, it does not appear to do so in Morgan-Short et al. (2012).

To summarize, the findings of the current experiment suggest that processing for form must be minimally conducted at the depth of interpreting, while simultaneously processing for
meaning, to potentially have a detrimental impact on comprehension. This negative impact can be explained by VanPatten’s (2004) Primacy of Meaning Principle, in that participants are limited in terms of cognitive resources, and processing for form at the greatest depth strains these resources to such a degree that processing for meaning can only be conducted at a subpar level. The current study found no difference between processing for form at the depth of identifying and doing so at the depth of interpreting. This result counters Morgan-Short et al.’s (2012) finding that depth of processing for form is positively related to comprehension, a difference that might be explained by the lack of randomization of depth of processing levels in Morgan-Short et al. (2012). In addition, Processing for Form at the Depth of Identifying does not appear to strain cognitive resources, and can be performed simultaneous to processing for meaning, without impacting the latter, at least in terms of comprehension. This result contrasts with VanPatten (1990), Greenslade et al. (1999), and Wong (2001). This divergence might be explained by the current study’s implementation of measures to control for processing for meaning and use of a multiple-choice test for comprehension.

Research Question II

The second research question, viz., Does Reactivity (TANTA) have an impact on L2 comprehension?, can also be answered affirmatively. The results of the ANOVA revealed that participants that thought aloud outperformed participants that did not in terms of multiple-choice reading comprehension with a large effect size ($\eta^2 = .145$; following Cohen, 1988). This result contradicts the expectations of the current researcher, given the conceptualization of thinking aloud as a cognitively demanding task that would theoretically have a negative impact on the main task being performed, and also contradicts the findings of previous studies on reactivity for
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

comprehension (Bowles & Leow, 2005; Goo, 2010; Leow & Morgan-Short, 2004; Morgan-Short et al., 2012; Rossomondo, 2007). Nonetheless, major differences can be pointed out between these previous studies and the current study, in terms of the Think-Aloud practice instructions and follow-up practice Think-Aloud task (Bowles & Leow, 2005; Leow & Morgan-Short, 2004; Morgan-Short et al., 2012; Yoshida, 2008), participant proficiency level (Bowles & Leow, 2005), the reading task (Goo, 2010; Leow & Morgan-Short, 2004), and the assessment task (Rossomondo, 2007).

Leow & Morgan-Short (2004), Morgan-Short et al. (2012), and, with regard to the non-metacognitive Think-Aloud group, Bowles and Leow (2005), gave participants fewer explicit instructions. In Leow & Morgan-Short (2004), participants were simply asked to “clearly speak their thoughts aloud throughout the entire experiment” (p. 46), similar to the instructions given in Morgan-Short et al. (2012), and in Bowles and Leow (2005) for the non-metalinguistic condition. Goo’s (2010) instructions, used with very minor modifications for the pilot study of the current experiment, were more detailed, but still did not sufficiently clarify the expectations for participants. As pointed out in Rossomondo (2007), instructions on how to think aloud should be careful to “avoid any misunderstandings about the researchers’ expectations,” with such potential misunderstandings possibly leading the participants to “feel [i] silly or embarrassed about externalizing their thoughts while reading” (p. 51). Indeed, Leow (2001), a study that used similar Think-Aloud instructions to Leow and Morgan-Short (2004), notes that, while not the predominant strategy, that reading the text in Spanish only was one that was employed by participants. This finding is corroborated by the pilot study of the current dissertation, which employed similar instructions to those used in Goo (2010), and found, out of a total of 8 Think-Aloud participants, that one participant simply read the text aloud in Spanish as opposed to
thinking aloud his thoughts, and two participants were reading the text aloud in Spanish until they were reminded that they needed to comment. This compares to the current study where no participant was found to only read the text in Spanish. One might speculate then that differences in the think-aloud instructions and the think-aloud practice task may have had some impact on some participants’ think-aloud protocols.

Key differences can be pointed out in terms of the reading tasks in Leow & Morgan-Short (2004) and Goo (2010), and the reading task of the current study. First, unlike the current study, Leow and Morgan-Short (2004) and Goo (2010) alleviated the burden of difficult vocabulary for students. In Leow and Morgan-Short (2004), participants were given “a take-home worksheet composed of two vocabulary matching tasks” (p. 45), and in Goo (2010) participants were provided with a glossary. The measures taken in both cases arguably facilitated the reading task for participants in the two studies. Indeed, Leow and Morgan-Short (2004) note that these tasks were purposefully “designed to reduce the processing load of content words” (p. 45). Furthermore, compared to the text of the current study, Leow and Morgan-Short (2004) and Goo’s (2010) reading task are arguably less difficult in terms of its morphology and content. Leow and Morgan-Short’s (2004) text is written entirely in the present tense and the imperative, does not include object pronouns, and reflects relatively common knowledge about healthy living, such as the need to wear a helmet when riding a motorcycle. Similarly, the text used in Goo (2010) is written entirely in the present tense, almost entirely in the first person plural, and is written in simple prose. The context of Goo (2010), about going on vacation, is also likely to be familiar to students, and includes such phrases as “In the canyon it is very hot and sunny and we are going to need a hat and sunscreen” (my English translation from the Spanish, p. 749). This compares to the text of the current study, which uses the present, imperfect, and preterit
tenses, employs direct and indirect object pronouns, fairly extensive prose, and is about the Aztecs. Moreover, participants in the current study were not given any help with vocabulary, requiring them to depend on the context. While one might point out that the average comprehension score in Leow & Morgan-Short (2004) is similar to that of the current study, a direct comparison cannot be made given that the assessment tasks were different in that Leow & Morgan-Short (2004) included open-response questions. The impact of thinking aloud while reading a more difficult text, like the one used in the current study, is arguably much more likely to be felt, which may explain the difference in results.

A similar argument can be made with regard to the neutral impact of the non-metacognitive concurrent verbal report in Bowles & Leow (2005). Bowles & Leow (2005), examined the impact of thinking aloud on a much more proficient group of participants, namely, fifth-semester students of Spanish. Due to their higher proficiency compared to the second and third-semester students of the current study, these participants may have been less likely to have difficulty with comprehending the text, making thinking aloud less likely to be helpful.

While similar in terms of the Think-Aloud instructions employed, Rossomondo (2007) and the current study differ with regard to the comprehension assessment task employed. In her Multiple-Choice Comprehension Task, Rossomondo (2007) included 13 questions with four potential answers, which varied only in terms of whether they were in the present, past, present perfect, or future tense. The correct answer for all 13 questions was the answer in the future tense. Given that the correct answer did not require the participant to understand the text, except to notice that it was written in the future tense, this particular task is limited in the degree to which it measured comprehension. In this way, the assessment task may have made it more difficult for thinking aloud to have an impact on the results, as it seems unlikely that thinking
aloud would be more beneficial to determining the tense of the text. This compares to the multiple-choice assessment task of the current experiment in which participants answered questions based on a variety of details about the text that they had read, and therefore, performance was dependent on comprehension.

As explored earlier in the literature review of the current dissertation, Morgan-Short et al. (2012) did not allow for a fair comparison between the Think-Aloud and the Non-Think-Aloud groups in their study. In the study, participants in the Think-Aloud group, but not in the Non-Think-Aloud group, were eliminated for rereading in order to ensure that participants were processing for form and meaning simultaneously. One might argue, however, that rereading may be helpful to comprehension, and therefore, eliminating participants that did so in one group but not the other may have had a confounding effect on the results. The current study, taking note of this possibility, as well as Morgan-Short et al.’s (2012) suggestion to instruct participants not to reread (p. 681), included explicit video instructions that specifically directed participants not to “go through the text for meaning and then go back and click on the forms, or click on the forms and then go back and read the text for meaning, [because] these two activities, clicking to label verb forms and reading for meaning need to occur simultaneously” (see Appendix I). In addition, the current study included a countdown time pressure, as well as a countdown timer, to discourage participants from rereading and/or deliberately taking additional time to compensate for the additional burden of processing for form. Probably due to these two improvements in the current study, only one already eliminated participant was found to have returned to the beginning to reread a substantial portion of the passage in the screen recordings or in the Think-Aloud verbal protocols. Therefore, the current study allowed for a fair comparison between the
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING THOUGHT, AND STRATEGIC PROCESSING

Think-Aloud and Non-Think-Aloud groups, which likely accounts for the difference in results between the current study and Morgan-Short et al. (2012) with regard to reactivity.

The finding of positive reactivity in the current study negates Ericsson and Simon’s (1984, 1996) theory on the validity of Think-Alouds. As explored in the literature review, Ericsson and Simon (1984, 1996) postulated three types of verbalization, one of which, Type II, is relevant to the current discussion. Ericsson and Simon (1984) defined Type 2 verbalization as that which requires participants to recode their thoughts, but not to explain their thought processes. The current study, however, employed this type of verbal report and found evidence of a positive impact on performance.

It may be, however, that the positive reactivity found for non-metalinguistic concurrent verbal reports in the current study is due to the type of medium employed, namely, comprehension. Different from solving a math equation or a crossword puzzle, comprehension of a text cannot be described as a simple step-by-step process. Strategies included reading sentences or sections of the passage and subsequently translating them into English, processing the text simultaneously in Spanish and English, and making comments. It may be that thinking aloud invigorated these strategies and motivated participants to put all their effort into the activity, potentially processing for meaning at a greater depth, which, following Craik and Lockhart (1972), improves memory of the processed material. Indeed, in a post-hoc analysis, participants in the Think-Aloud group did rate their effort to be significantly higher than the Non-Think-Aloud group. Another potential explanation is that employing one’s strategies was simply more effective when done out loud than silently. One might speculate that this aided participants’ recognition of specific details of the text, such as dates and colors, or alternatively, that participants were better able to assess and improve their strategies as they progressed through the
reading because they were able to hear them audibly. While one might point out that Leow and Morgan-Short (2004) also found evidence of participants using these types of strategies, given that their reading task was arguably less difficult, particularly in terms of vocabulary, it may be that the impact of these strategies was less notable. Regardless of the reason(s) behind the positive reactivity found in the current study, which remain speculation at this point, the results of the current study pertain to reading comprehension, and should not be extrapolated to other mediums.

Research Question III

The third research question, viz., *Is there an interaction between Depth of Processing for form and Reactivity (TANTA) with regard to their impact on L2 comprehension score?*, cannot be answered in the affirmative, as no main effect was found for interaction in the ANOVA performed. Interestingly, this indicates that the positive impact of thinking aloud was no more helpful to participants in the detrimental Processing for Meaning and Form at the Depth of Interpreting Condition than in the Processing for Meaning Only group. This appears to contradict the idea expressed in the discussion of Research Question II, namely, that thinking aloud might be particularly beneficial to participants that are pressed for cognitive resources by the additional task of processing for form. Nonetheless, this somewhat unexpected result might be the result of the elimination of participants that exceeded the time-limit by more than 15 seconds. Six participants in both the Think-Aloud group and the Processing for Meaning and Form at the Depth of Interpreting were eliminated for this reason, a number that equaled that of all the participants that did so in all of the remaining groups combined. Given that the average score of these six participants was higher (59.17) than that of non-eliminated participants in both the
PROCESSING for Form at the Level of Interpreting condition and the Think-Aloud group (55.35), it is possible that those participants that benefited most from the think-aloud condition may have been eliminated, thus reducing the likelihood of an interaction between depth of processing for form and TANTA. It should also be noted that the effect size was small ($\eta^2 = .02$; following Cohen, 1988), indicating that the number of participants might have played a role.

Research Question IV

The forth research question, namely, *Is there a relationship between amount (number of clicks) of processing for form and L2 comprehension score?*, cannot be answered in the affirmative. In light of the negative impact of processing for form at the depth of interpreting while simultaneously processing for meaning, this result seems to indicate that depth of processing, as opposed to amount of processing, may play the critical role. Nonetheless, given that participants were eliminated if they did not process at least 60% of the forms, and, in the case of participants assigned to the depth of interpreting, at least 75% of those at the proper depth, it may be that amount of processing also plays a role, but that once a minimum amount of processing is reached, more processing for form does not have an additional impact.

This explanation is consistent with the fact that Morgan-Short et al. (2012) found depth of processing for form, excluding the processing for meaning only group, to be positively related to comprehension. In their study, while participants were eliminated if they did not process a minimum number of forms, participants only needed to process one form at a higher depth of processing to be coded for that level. This contrasts with the current study, which required that participants process at least 75% of the forms at the depth of interpreting to not be eliminated, and found no significant difference between processing for form at the depth of identifying and doing so at the depth of interpreting. One might speculate that the positive relationship between
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

depth of processing and comprehension in Morgan-Short et al. (2012) may have been partially
due to the fact that participants were not consistently processing at the higher depths they were
coded for, and therefore, were unlikely to affected by the potentially neutralizing impact of
processing for form at higher levels of processing. Indeed, as noted in the literature review, while
Morgan-Short et al. (2012) did not include precise figures with regard to the number of forms
that participants coded as processing at the depth of interpreting typically processed at that level,
Leow et al. (2008), whom Morgan-Short et al. (2012) conceptually replicated, found that “the
majority of participants who translated or interpreted the target forms did so on very few
items…” (p. 681).

Research Question V

The fifth research question, namely, *Is Working Memory Capacity related to L2 Comprehension?*, cannot be answered in the affirmative, as Working Memory Capacity was not
found to be related to Comprehension. This result is contrary to previous research which has
largely found working memory to be related to comprehension (e.g., Daneman & Carpenter,
1980; Engle et al, 1992; Harrington & Sawyer, 1992; Leeser, 2007; Walter, 2004), although not
always (e.g., Chun & Payne, 2004; Goo, 2010). While the current study does differentiate itself
from those studies that found such a relationship in that it included while-reading tasks, it seems
unlikely that working memory capacity would play less of a role due to the inclusion of
additional tasks. Indeed, as no interaction was found between working memory capacity and
reactivity or between working memory capacity and depth of processing, an argument based on
these additional tasks does not seem to hold up.

The current study, however, can be seen as distinct to previous research finding evidence
of a relationship between working memory capacity and reading comprehension when these
studies are examined on an individual basis (Daneman & Carpenter, 1980; Engle et al, 1992; Harrington & Sawyer, 1992; Leeser, 2007; Walter, 2004). First, of these studies, only Harrington and Sawyer (1992), Lesser (2007) and Walter (2004) examined L2 comprehension. This is crucial to point out, as it may be less likely for the relationship of working memory capacity to be detected in the second language given the potentially greater impact of varying degrees of L2 vocabulary. Among these three studies (Harrington & Sawyer, 1992; Leeser, 2007; Walter, 2004), Leeser (2007) and Walter (2004) measured comprehension via a summary completion task and a comprehension recall post-test. As they require the participant to recall information, these tasks are arguably more reliant on the storage component of working memory capacity than the multiple-choice task employed in the current experiment, and therefore, they may have had an influence on the finding of a significant relationship between working memory capacity and reading comprehension in the two studies. It might be pointed out that Goo (2010) also used a measurement for comprehension that was dependent on recall, namely, giving participants a series of questions to answer without multiple-choice options, and did not find evidence of a relationship between working memory capacity and comprehension. As Goo (2010) notes, however, the Think-Aloud condition, markedly different than in the current study in that participants were not provided with an example of how to think aloud, may have played a role. Harrington and Sawyer (1992), like the current study, relied upon participants’ scores on the Verbal Scholastic Aptitude Test, which is a multiple-choice test. This particular test, however, allows participants to backtrack, unlike in the current experiment where participants were electronically blocked from doing so. This may have increased the dependence of participants on the processing component of working memory capacity, as participants most likely referenced the passage while they answered questions. While the current experiment did involve additional
processing during the reading task for participants in the processing for meaning and form groups, the Multiple-Choice Comprehension Task involved only minimal processing, as participants merely needed to answer the questions. In addition, participants were likely more motivated during the Verbal Scholastic Aptitude Test, given that it had an impact on their future. This motivation may have led participants to maximize the use of their processing resources, potentially making working memory capacity differences more notable. In summary, studies finding a relationship between L2 reading comprehension and working memory capacity have either included a higher degree of processing (Harrington & Sawyer, 1992), in terms of the reading task, or storage (Lesser, 2007; Walter, 2004), in terms of the assessment tasks, which, as the two crucial components of working memory capacity, likely had an impact on the findings.

The lack of a finding for a relationship between working memory capacity and L2 comprehension found in the current study suggests that the generally accepted relationship between working memory capacity and reading comprehension may need to be moderated. It may be, given the significant relationships found in L2 studies with particularly strenuous processing and storage components (Harrington & Sawyer, 1992; Lesser, 2007; Walter, 2004), that working memory capacity may be related to aspects of reading comprehension tasks, as opposed to reading comprehension in general.

It is also possible that the mathematical nature of the operation span task may have played a role in the results, a result that is further explored in the discussion of the following research question.
Research Question VI

The sixth research question, namely, *Is there any interaction between Working Memory Capacity and Depth of Processing for form in their relationship to L2 comprehension score?*, also cannot be answered in the affirmative. This indicates that individuals with more cognitive resources are not better able to process for form and meaning simultaneously, meaning, contrary to expectations, that working memory capacity was not a mediator variable between Comprehension and Depth of Processing. The finding contradicts studies that have found that working memory capacity is related to multitasking (Bühner et al., 2006; Colom et al., 2010; König et al., 2005; Salthouse & Miles, 2002). The result is particularly surprising given that König et al.’s (2005) two tasks included marking identical numbers across two columns and reading and answering questions based on a filled-out calendar, which are conceptually similar to the two tasks of the current experiment, clicking on forms and reading a text. One might speculate, however, that two key differences between the current study and König et al. (2005) may explain the results. One notable difference between the current study and König et al. (2005) is that the current study involved a foreign language. It is possible that differences in vocabulary or language aptitude may have played a more important role in the current study, making the relationship of working memory capacity more difficult to detect. Another notable difference is that in König et al. (2005) participants answered questions while they read the calendar, as opposed to after, and therefore, likely required more processing resources, making working memory capacity more crucial to success.

The finding that working memory capacity is not related to the ability to simultaneously process for form and meaning also seemingly contradicts VanPatten’s (2004) Input Processing Principles, whose foundation is predicated on the notion that limits in cognitive resources restrict
the degree to which participants can process different aspects of input, which implies that individuals with more available resources should be better able to process for multiple aspects of input, such as form and meaning. Nonetheless, the processes involved in an operation working memory capacity task are arguably different from those involved in reading a text for meaning and form. First, the working memory capacity task required participants to recall words, and therefore involved a strong memory component. While it might be argued that participants in the reading comprehension task also needed to remember what had occurred in the passage, participants only needed to recognize, not recall, the correct answer, as noted in the discussion of the previous research question. Additionally, the processing component of the working memory capacity task required participants to make mathematical calculations, whereas processing for form involved identifying and/or interpreting forms. One might speculate that individuals who are talented at solving mathematical equations might be less so at language tasks, or, more specifically, at differentiating between the imperfect and the preterit aspects. The lack of a relationship between depth of processing for form and working memory capacity may be, therefore, due to the confounding effect of participant abilities.

Alternatively, it may be that processing for form, rather than occupying additional cognitive resources, is distracting to participants, and indeed, many participants noted in the Post-Test Survey that they found it difficult to process for meaning and form simultaneously. While it might be noted that the ability to deal with distractions has been found to be related to working memory capacity in cognitive research (Kane & Engle, 2000; Kane, Bleckley, Conway, & Engle, 2001), the distraction conditions in these studies included finger tapping and an antisaccade task, each of which is arguably quite different from processing for past forms.
Research Question VII

The seventh research question, namely, *Is there any interaction between Reactivity and Working Memory Capacity in their relationship to L2 comprehension score?*, as in the case of the previous two research questions regarding working memory capacity, cannot be answered in the affirmative. This suggests that WMC was not a mediator variable between Reactivity and Comprehension. As articulated in the discussion of Research Question II, the expectation of the impact of thinking aloud was that it would negatively impact comprehension, given that learners would potentially need to devote cognitive resources to thinking aloud that would otherwise be able to be devoted to comprehension. Therefore, it was expected that working memory capacity would potentially alleviate the postulated negative impact of thinking aloud on comprehension. In light of the positive impact that reactivity was found to have on comprehension, however, the finding that reactivity and working memory capacity do not interact in their relationship to L2 comprehension score is not particularly surprising, as given that thinking aloud was not burdensome in this experiment there is no expected advantage for individuals with more cognitive resources.

Pedagogical Implications

The findings of the current dissertation have pedagogical implications both with regard to the types of tasks to assign to students, as well as to the potential strategies students might employ to improve their comprehension. First, the significant negative impact of processing for meaning and form at the depth of interpreting suggests that instructors might consider limiting the focus of reading activities, or alternatively, asking students to only focus on one aspect of a text at a time. For example, the instructor might consider giving students a text and ask them to read for comprehension, and subsequently, ask them questions about the meaning of the text.
Then, the instructor could ask participants to go back and reread the text, but this time he could ask them to focus on a grammatical aspect of the text, such as the context of the use of the imperfect vs. the preterit (cf. Leow, 2009, pp. 28-29 for a similar pedagogical extrapolation).

Second, the positive impact of thinking aloud on comprehension and the qualitative finding that translation was a strategy common to all the think-aloud participants, suggests that instructors may want to devote time to developing reading strategies. Specifically, instructors may want to encourage students to avoid reading the text directly in the foreign language, and attempt to comment and translate as they proceed through the text. When texts are difficult given the proficiency of students, they may want to encourage students to do so aloud.

Limitations

The current study includes several potential limitations. First, the processing aspect of the working memory capacity task, solving mathematical equations, may have had an impact on the results, given the potentially confounding factor of mathematical ability. Second, the current experiment included participants with languages profiles that included knowledge of other languages and native languages other than English, and individuals older than 22. Nonetheless, it should be kept in mind that, taken as a whole, the Multiple-Choice comprehension score of the former group (56.31) was relatively similar to the overall average (53.37). In addition, the scores of the two participants older than 22 was within the range of scores for their respective conditions. Third, the total number of participants (80) was relatively low for the analyses involving interaction. This is particularly evident given the low effect size of the non-significant result of Research Question III. Fourth, attrition was much higher for those participants in both the Think-Aloud group and the Processing for Meaning and Form at the Depth of Interpreting.
Condition, as the number of participants eliminated for taking too much time was equivalent to the remaining five groups combined. This may be a concern given that the average score among participants who took too much time was higher than that of non-eliminated participants (60.83 vs. 55.37). This being said, the counterargument is that these individuals may not have been proceeding as quickly as possible, and, in this way, were not following the instructions of the experiment. A final potential limitation of the current experiment is that it involved reading comprehension, and not listening comprehension, and therefore, cannot be directly compared to the results of experiments on processing for form and meaning conducted on aural comprehension (VanPatten, 1990; Wong, 2001).

Future Research

Future research should address the limitations of the current study. Most importantly, future research should consider using a larger initial group of participants, in order to allow for eliminations based on language profiles. Future research should also consider employing multiple working memory capacity tests to determine working memory capacity, in order to reduce the potential impact of mathematical ability. Future research might also consider expanding on the current study to address, for example, modality, level of proficiency, type of linguistic item, type of assessment tasks, and so on.

With regard to the impact of depth of processing, future research might consider conceptually replicating the current study in the aural mode. Processing for form and meaning has the potential to have an even greater impact in the context of the additional task of decoding the audio. It would also be worth exploring the impact of processing for form and meaning for participants with higher proficiency levels. The expectation would be that processing for form
and meaning simultaneously would have less of an impact at this level, as a greater vocabulary, among other reading-related skills, would lessen the amount of processing required to comprehend the text. One might also explore the ability of participants to focus on pronunciation and meaning simultaneously, in order to determine if focusing on aspects other than grammar also has an impact on comprehension. Future research could also look at the relationship between working memory capacity and depth of processing in terms of their impact on form production and recognition.

Working memory capacity is also deserving of future investigation in the context of processing for multiple aspects of a reading. In light of the non-significant relationship between working memory capacity and comprehension in the current study, for example, future research may want to explore the use of working memory capacity tasks with processing components that minimize the impact of participant abilities not related to cognitive resources. For the same reason, it would be worthwhile to examine the impact of type of comprehension test (e.g., recall vs. multiple-choice) on the possible relationship between working memory capacity and comprehension. As articulated in the discussion, it may be that the arguably more memory dependent comprehension test, namely, recall, is more likely to be related to working memory capacity. Future research might also explore the relationship between working memory capacity and depth of processing in terms of their impact on form production and recognition.

With regard to reactivity, several avenues of potential future research can be identified. First, one might compare the impact of video Think-Aloud instructions that include an example of thinking aloud, such as in the current experiment, to more typical written Think-Aloud instructions. Such a study could also explore proportional differences in strategies employed by participants in the two distinct instructional groups. As noted in the discussion section, it is likely
that including an example of thinking aloud in the instructions results in participants more consistently thinking aloud, as opposed to simply reading the text in the foreign language. Finally, it would be of interest to Second Language Acquisition to qualitatively analyze Think-Aloud protocols to investigate the possibility that certain strategies are related to improved performance.

Conclusion

The current experiment found partial evidence to support VanPatten’s (2004) Primacy of Meaning Principle. Participants that processed simultaneously for meaning and form at the depth of interpreting were found to differ significantly from their counterparts in the processing for meaning only group. No such difference, however, was found between participants processing for form at the depth of identifying and participants processing only for meaning, or between participants processing for form at the depth of identifying and those doing so at the depth of interpreting. The former contradicts Greenslade et al. (1999), VanPatten (1990), and Wong (2001), who found evidence that processing for meaning and form at the depth of identifying had a negative impact on comprehension, and supports the lack of a finding of such an impact in both Leow et al. (2008) and Morgan-Short et al. (2012). The lack of a significant difference between the two processing for form conditions, however, refutes Morgan-Short et al. (2012), who found depth of processing for form to be positively related to comprehension. Methodological differences between the studies, such as coding for depth of processing a priori, may have played a role in the difference in results.

The current experiment also found evidence for positive reactivity, contrary to previous research on reactivity for reading comprehension (Bowles & Leow, 2005; Leow & Morgan-
The current experiment did not find a relationship between reactivity and depth of processing. While this finding is somewhat surprising in light of the positive impact of thinking aloud found in the current study, it is possible that those participants that were most helped were eliminated by the time-limit.

No relationship was found between amount of processing and comprehension. This may indicate that depth of processing, rather than amount of processing, is the major factor impacting comprehension. Alternatively, given that participants were eliminated for not processing a sufficient number of forms, it may be that a ceiling effect had been reached in terms of the negative impact of processing for form.

Interestingly, the current experiment did not find evidence for a relationship between working memory capacity and comprehension, contrary to previous research finding evidence of such a relationship (e.g., Daneman & Carpenter, 1980; Engle et al, 1992; Harrington & Sawyer, 1992; Leeser, 2007; Walter, 2004). This might be due to the arguably higher processing and storage components of the reading comprehension task in previous experiments, which may have made them more favorable to higher working memory capacity participants. Therefore, the widely accepted relationship between working memory capacity and reading comprehension may need to be moderated, particularly as it pertains to the L2. Alternatively, mathematical ability may have been a confounding factor for the working memory capacity task, a possibility...
that is discussed further in regard to the lack of an interaction between working memory capacity and depth of processing (below).

The current experiment also did not find evidence of an interaction between working memory capacity and depth of processing in terms of their relationship to comprehension. This result is somewhat surprising given that working memory capacity is assumed to be linked to available processing resources. Nevertheless, the use of a multiple-choice comprehension task may have limited the importance of the storage component of working memory capacity, as participants merely needed to recognize the correct answer. Moreover, the processing component of the current task was grammatical in nature, compared to the mathematical nature of the operation span working memory capacity task. It may be that the same participants that were mathematically inclined may have been less familiar with these conjugations, making any interaction between working memory capacity and depth of processing in terms of their relationship to comprehension less likely to be detected.

Lastly, no interaction was found between reactivity and working memory capacity in terms of their relationship to comprehension. In light of the finding that the Think-Aloud condition was beneficial to performance, however, this result is not surprising, as working memory capacity would only have been expected to play a role in the event that thinking aloud strained processing resources.

---

i Lexical Temporal Indicators- abbreviated as LTI, these are cues that provide hints to the reader as to the tense
ii Gustar- meaning “to like,” this Spanish verb’s conjugation agrees with theme and its clitic agrees with the experiencer (Bowles 2008).
iii Due to substantial skipping by this participant, it was not possible to assign an exact WM score to him, which provided additional motivation for this particular participant to not be included in the statistical analyses.
iv Participants selected a percentage multiple of ten or zero (see Appendix D).
v The researcher thought this information would be helpful while in the process of conducting the pilot. For this reason, the period of time between the second session and the answering of this question varied.
vi During one slot of the second session, the server of the computer program was down, and participants in this group participated in the second session up to one week later.
There is a 3.8% chance of a student correctly labeling 12 out of the 16 (75%) past forms by chance. Comparatively, there is a 1.1% chance of a student correctly labeling 13 out of the 16 forms by chance. As 5% is generally established as the standard acceptable rate of error in SLA (e.g., alpha value in Wong, 2001; Morgan-Short et al, 2012), the 3.8% chance was determined to be acceptable.

A line for the last WMC recall set was accidentally erased by the researcher, and neither this recall set nor its corresponding processing set were included in the analysis.

\[ t(13) = 1.104, p = 0.353. \]

The fourth research question was added after the pilot.

One of the three participants that was not part of the final analysis also simply read the text aloud.

This participant was the participant eliminated for not processing for meaning.

While data from one of these participants was not included in the final statistical analysis because the participant signed up late for the study, data from this participant is nonetheless relevant to the discussion on the think-aloud instructions and practice task.

As noted earlier, only one of these participants was included in the final analysis. However, his inclusion in this particular case, namely, whether to change the Think-Aloud instructions, was deemed relevant.

Ages were 29 and 41.

Instances of hacia and sabiduría were not counted towards potential elimination for clicking on too many target forms given their close resemblance to the imperfect.

There is a 3.8% chance of a student correctly labeling 12 out of the 16 (75%) past forms by chance. Comparatively, there is a 1.1% chance of a student correctly labeling 13 out of the 16 forms by chance. As 5% is generally established as the standard acceptable rate of error in SLA (e.g., alpha value in Wong, 2001; Morgan-Short et al, 2012), the 3.8% chance was determined to be acceptable.

For two participants, the screen recordings were not saved properly. The Post-Test Survey and the time between clicks, however, indicated that the participant had processed for meaning.

Due to computer issues with one participant, the two sets that were displayed during the computer issues were omitted for this participant.
Sample Portion of Text

Para mí, Frida Kahlo es una de las más grandes artistas mexicanas. Cuando tenía 5 años sufrió una enfermedad muy grave llamada polio, y Frida sobrevivió, pero una de sus piernas le quedó muy débil, más corta y más delgada que la otra. Su padre era fotógrafo y artista y se cree que mantenía una relación muy profunda con Frida.
Los Aztecas (Rangel Montemayos)

El pueblo azteca, como pueblo primitivo, descubrió una solución a los problemas presentados por las fuerzas de la naturaleza. Se sabe que el pueblo azteca daba mucha importancia a su religión, sobre todo a su dios principal y todopoderoso Tonatiuh (el sol). Tonatiuh poseía las mismas bondades y los defectos de los humanos, pero con un gran poder sobrenatural. Durante su historia, los antiguos mexicanos le sacrificaron muchos humanos al sol pero no por crueldad ni instintos bárbaros, sino por respeto y adoración.

Los Aztecas construyeron muchos monumentos para honrar al sol que tanto admiraban. Entre estos monumentos el más importante era la Piedra del Sol. Hoy día, se la conoce también con los nombres de Calendario Azteca o Jícara de Águilas (el pueblo Azteca le dio el nombre de Cuauhxicalli). El Calendario Azteca es una de las obras de arte más hermosas de esta cultura. Es un monolito o monumento de piedra.
APPENDIX C

Brief Questionnaire

1. How old are you? ____________________________

2. What languages do you speak?

2a. Since I was between 0 and 6 years old, I have spoken ____________________.

2b. Since I was between 6 and 12, I have spoken ____________________.

3. What languages have you studied?

3a. I studied ________________ from age ________________ to ________________

   for a total of ________________ years. I believe my proficiency in this language is (circle one)

   BEGINNER          INTERMEDIATE         ADVANCED

3b. I also studied ________________ from age ________________ to ________________

   for a total of ________________ years. I believe my proficiency in this language is (circle one)

   BEGINNER          INTERMEDIATE         ADVANCED

3c. I also studied ________________ from age ________________ to ________________

   for a total of ________________ years. I believe my proficiency in this language is (circle one)

   BEGINNER          INTERMEDIATE         ADVANCED

4. Do you have any known visual or hearing impairments? (which eyeglasses or contact lenses are not enough to fix). Circle one of the following statements:

   [ ]

239
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

**NO** I do not have any visual and/or hearing impairments.

**YES** I have a visual and/or hearing impairment. Explain if YES: ______________

________________________________________________________________________
APPENDIX D

Post-Test Survey

What percentage of the content you just read did you know PRIOR to reading the article?

0%  10%  20%  30%  40%  50%  60%  70%  80%  90%  100%

What percentage of the comprehension questions did the knowledge you had PRIOR to reading the Aztecs article help you to answer?

0%  10%  20%  30%  40%  50%  60%  70%  80%  90%  100%

Please rate your effort on the activities you completed during today’s session on a scale of 1 to 5.

1  2  3  4  5
Low  High

While reading the Aztecs text, do you feel you were able to understand the meaning?

1 (No)  2  3  4  5 (Yes)

In 2-3 sentences, please explain.
APPENDIX E

Sample Selection of Multiple-Choice Practice Task

1. When Frida Kahlo was five years old, she suffered from a disease known as
   a. Tuberculosis.
   b. Polio.
   c. Tetanus.
   d. Malaria.

2. She survived, but
   a. she lost her sense of hearing
   b. one of her arms were left shortened.
   c. one of her legs were left shortened.
   d. two of her fingers were left thinner.
Sample Selection of Multiple-Choice Comprehension Task

1. Tonatiuh was the Aztec’s solution to
   a. win battles.
   b. unexplicable deaths.
   c. the forces of nature.
   d. find peace.

2. Tonatiuh possessed the following characteristics:
   a. a great supernatural power.
   b. human strengths, but accompanied by a supernatural power
   c. human weaknesses, but accompanied by a supernatural power
   d. human strengths and weaknesses, together with a supernatural power.

3. The Aztecs performed human sacrifices
   a. to appease Tonatiuh.
   b. to appease their gods and also because they were barbaric warriors.
   c. because they were known for their barbaric instincts and cruelty
   d. all of the above

4. The Aztecs are the ancestors of
   a. Mexicans
   b. Peruvians
   c. Argentinians
   d. Cubans

5. The Aztecs constructed
   a. many monuments to honor Tonatiuh.
   b. some monuments to honor Tonatiuh.
c. only one monument to honor Tonatiuh.

d. three monuments to honor Tonatiuh.

6. The most important monument in Aztec life was

   a. Jícara de Aguilas.
   
   b. Piedra del Sol.
   
   c. Cuauhxicalli.
   
   d. All of the above.

7. The Calendario Azteca is one of the

   a. largest works of the Aztec culture.
   
   b. most beautiful works of the Aztec culture.
   
   c. oldest works of the Aztec culture.
   
   d. None of the above.

8. The Calendario Azteca was

   a. the famous Aztec calendar known for its accuracy.
   
   b. the name of a well-known museum in Mexico.
   
   c. a monument made of stone
   
   d. a special ritual for Tonatiuh.
APPENDIX G

The following is an explanation of the TA video instructions.

[Audio and Visual]

In this experiment, I want to hear your thoughts, so I want you to think your thoughts aloud throughout the experiment into the microphone. You need to think aloud starting with the Think-Aloud practice task and then throughout the experiment. That means during the reading and during the questions about the reading, until the end of the experiment. In other words, once you start, don’t stop thinking aloud. Don’t plan what to say, just let your thoughts speak for themselves. The following is an example of a Think-Aloud protocol.

[Visual]

Dos palabras

Tenía el nombre de Belisa Crepusculario, pero no por fe de bautismo o acierto de su madre, sino porque ella misma lo buscó hasta encontrarlo y se vistió con él. Su oficio era vender palabras.

[Audio]

“hmm… two words… so Belisa was her name… pero no por fe de bautismo… so she wasn’t baptised with this name… or her mother’s wishes? sino porque she misma lo buscó hasta encontrarlo… and she wore her name… her office, no, job? was to sell words.

[Audio and Visual]

I would like you to talk CONSTANTLY, CLEARLY, and LOUDLY. Please take your time.

Please practice thinking aloud with the following task.

Think aloud, that is, verbalize aloud or say whatever comes to your mind while you perform this task. You don’t need to explain what you are saying just talk aloud as you do the task: Read the following text, from the same short-story “Dos palabras,” for comprehension.

[Visual]

Vendía a precios justos. Por cinco centavos entregaba versos de memoria, por siete mejoraba la calidad de los sueños, por nueve escribía cartas de enamorados, por doce inventaba insultos para enemigos irreconciliables. También vendía cuentos, pero no eran cuentos de fantasía, sino largas historias verdaderas que recitaba de corrido sin saltarse nada.
APPENDIX H

Post-Test Survey

What percentage of the content you just read (the Aztecs reading) did you know PRIOR to reading the article?

0%  10%  20%  30%  40%  50%  60%  70%  80%  90%  100%

What percentage of the comprehension questions did the knowledge you had PRIOR to reading the Aztecs article help you to answer?

0%  10%  20%  30%  40%  50%  60%  70%  80%  90%  100%

Please rate your effort on the Aztecs reading and the questions associated with this reading on a scale of 1 to 7.

1  2  3  4  5  6  7

Very Low

Very High

On a scale of 1 (very minimal) to 7 (very hard), to what degree did you, as you read the Aztecs reading, try to understand the meaning of the text?

1  2  3  4  5  6  7

(Very Minimal)  

(Very Hard)

In 2-3 sentences, please explain.

[open text box]

[For conditions 2 and 3: Besides the past tense forms,] what did you focus on during the Aztecs reading?

[open text box]
APPENDIX I

The following is an explanation of the Condition 3 video instructions.

[START VIDEO 2]

[Audio]

When you see these BOXES, there’s are going to be boxes that ask you to please wait until I tell you to proceed, and there are going to be boxes that are just going to ask you a question.

[Students will visualize an example of one of these boxes]

[Visual]

Please read as quickly as you can the following text on Frida Kahlo for comprehension. You will be asked to answer some questions after your reading without referring back to the text. At the same time, please click once on all instances of the imperfect and twice on all instances of the preterit as you read. You’ll see [IMP] appear after those that you have clicked once, indicating that you believe they are imperfect forms, and [PRET] appear after those you have clicked twice on, indicating that you believe they are preterit forms. You may click an additional (third time) to erase your marking. [Think-aloud group only] Please remember to think-aloud throughout the experiment. I will tap on your shoulder if you are silent or simply reading aloud as opposed to thinking aloud.

[Audio]

Now I’m going to be going over the instructions for reading the text to follow. You need to read the text for comprehension, meaning that you are paying attention to the meaning of the text, and you also need to, as you read, click on verb forms in the past. Click on imperfect verbs once. Click on preterit verbs twice. As you click on them, there is either going to be IMP in brackets that appears after the form or PRET in brackets next to the form meaning imperfect and preterit, respectively. Please do not go through the text for meaning and then go back and click on the forms, or click on the forms and then go back and read the text for meaning, these two activities, clicking to label verb forms and reading for meaning need to occur simultaneously. Also, do not reread or go back to earlier parts of the reading for any reason.

Please proceed as quickly as possible, but don’t abandon either clicking on the forms or reading the text for comprehension in order to gain additional speed.

Keep in mind you can click on verbs an additional (third) time to erase your marking.

Also, there will be a timer on top of the screen and it will follow you as you scroll down. Please try to finish before the timer is up.

Once you finish the reading, immediately proceed to the next page.
Please remember to read all instructions carefully and not to proceed unless you are sure you understand what is required.

[Think-aloud group only; Audio and Visual]

Please remember to Think-Aloud throughout the experiment. I will tap on your shoulder if you are silent or simply reading aloud as opposed to thinking aloud.

[Both TA and NTA; Audio]

Please ask me any questions you might have.
REFERENCES


PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING


PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING


Craik, F. I. M. (2002). Levels of processing: Past, present... and future? Memory, 10(5-6), 305–318. doi:10.1080/09658210244000135


PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING


doi:10.1177/009365091018005005


Dekeyser, R. M. (2001). Automaticity and automatization. In P. Robinson (Eds.), *Cognition and
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING


PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING


PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING


PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING


PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING


PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING


PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING

doi:10.1037/0882-7974.6.1.118


doi:10.1111/j.1467-9922.2009.00500.x


doi:10.1109/THFE.1964.231647
PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING


PUSHING FOR PROCESSING: THE ROLES OF DEPTH OF PROCESSING, WORKING


