

ABSTRACT

Title of Document:

ADULT READERS' CALIBRATION OF
WORD LEARNING

Meghan M. Parkinson, Doctor of Philosophy
2011

Directed By:

Professor Patricia A. Alexander, Department of
Human Development

The current study examined undergraduates' metacognitive processes during word learning, a crucial component of building representations of key concepts from text. Noticing the need to construct meaning for unknown words requires metacognitive monitoring. Constructing meanings for those words requires regulation of cognition.

Fukking (2005) provided a model for word learning, based on think aloud data that represented a series of metacognitive activities word learners engaged in when faced with an unknown word. The evaluation process within Fukking's (2005) model related to the judgments learners made about new word meanings and how accurate they believed those judgments to be. A specific aspect of metacognitive evaluation is *calibration*, or the accuracy with which learners assess their knowledge on a particular cognitive task (Glenberg & Epstein, 1985; Lichtenstein & Fischhoff, 1977). The current study more closely examined word learning and calibration, while addressing

some gaps in the literature and offering a model of influences on word learning to complement Fukkink's process model.

The current study sought to answer questions related to the following goals: 1. To determine the influence of several factors related to adult readers' word learning and calibration of word learning. 2. To assess empirical evidence relative to a model of reading skill, vocabulary knowledge, passage comprehension, and metacognitive evaluation related to word learning using methods that directly measure word learning and metacognitive evaluation. 3. To determine which text factors influenced the ease with which word learners could derive meaning while reading and evaluate their level of performance on a word knowledge test.

A measured variable path analysis showed a similar goodness of fit for both the incidental word learning condition and the intentional word learning condition. Prior word knowledge was found to be positively related to judgments of learning, but negatively related to calibration of word learning within the path model. Think-aloud data did not illuminate a connection between passage comprehension, strategic processing, and word learning. However, think-aloud data did reveal that students who decreased in performance from word knowledge pretest to posttest self-reported challenge while reading more frequently than other students. Finally, repeated-measures ANOVAs revealed differences in passage comprehension and JOLs between passages, prompting an analysis of specific text features underlying text difficulty that were not represented with a readability formula.

ADULT READERS' CALIBRATION OF WORD LEARNING

By

Meghan M. Parkinson

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor Patricia A. Alexander, Chair
Ann A. Battle, PhD
Professor Gregory R. Hancock
Associate Professor Min Wang
Professor Kathryn R. Wentzel

© Copyright by
Meghan M. Parkinson
2011

Acknowledgements

My heartfelt thanks go out to a diverse group of people who are my memories and inspiration for forging a scholarly path. Nobody I know truly loves reading more than my Grandfather, Elmer Amsden, and he has shared his enthusiasm for whatever he was reading with me since I was old enough to understand what he was saying (and probably before that!). My desire to be a researcher surely started thanks to the influence of his curiosity.

My parents, Nord and Stephanie Parkinson, supported my early investigations into reading and writing by driving me miles to the nearest library. When I went to college they were back in the car, selflessly driving me five hours each way between semesters.

Heather Haverback has been a steady guide since my first day as a graduate student. Her advice, as both an academic mentor and friend, has helped me navigate the up, downs, twists, and turns of the graduate school terrain. Without her I would have truly lost my way.

One of the most fortunate twists of my graduate career was being adopted into Patricia Alexander's mentorship. Her guidance will always be my model of how to be a successful scholar and nurturing mentor. When she adopted me I was lucky enough to become part of a large unruly lab group that has allowed me to participate in ponderings and debates that have drawn new destinations on my map of ideas.

Finally, my husband, Daniel Dinsmore has ensured that I am a thriving rather than just surviving the deadlines and endless to-do lists. His patience and nurturing

have been boundless. Daniel's passion for the research process continually reminds me of the wonder in being able to pursue my curiosity.

Table of Contents

CHAPTER 1: INTRODUCTION	1
Statement of the Problem.....	6
Purpose of the Study	9
Research Questions and Hypotheses	9
Research Question 1	9
Research Question 2	10
Research Question 3	10
Research Question 4	10
Research Question 5	11
Definition of Terms.....	11
Significance of the Study	13
CHAPTER 2: LITERATURE REVIEW	15
Word Learning	15
A Model for Word Learning.....	21
Metacognition	26
Research Paradigms	29
Calibration.....	31
Word Learning Variables Studied in Relation to Calibration.....	38
Person Factors	39
Prior Knowledge	39
Purpose and Interest.....	44
Text Factors	46
Narrative vs. Expository Text	46
Text Difficulty	48
Context.....	48
Word Factors.....	49
Research Gaps.....	54
CHAPTER 3: METHOD AND DATA ANALYSES	56
Pilot Study.....	56
Purpose.....	56
Participants.....	56
Measures and Procedure	57
Method	63
Participants.....	63
Measures	64
Nelson-Denny Reading Test	64
Word-knowledge Pretest.....	65
Narrative Passages	69
Judgment of Learning Scales	72
Word Knowledge Posttest.....	72
Confidence Scales	73
Think-aloud Protocols.....	73

Procedures.....	75
Data Analyses	76
CHAPTER 4: RESULTS AND DISCUSSION.....	82
Multigroup Analysis of Incidental and Intentional Word Learning	83
The Influence of Word-Knowledge Pretest Scores on JOLS and Calibration...89	
Processes and Strategies Related to Word Learning and Calibration of Word Learning.....	90
Strategic Processes Reported and Highest Word Knowledge Gains	90
Strategic Processes and Largest Decline in Word Knowledge	92
Strategic Processes and the Highest Levels of Calibration.....	93
Strategic Processes and the Lowest Levels of Calibration	94
Effect of Passage Difficulty on Main Idea and JOL Scores	96
Differences in Readers' Comprehension	96
Differences in Readers' Judgments of Learning.....	97
Text-level and Word-level Factors Affecting These Observed Differences.....	98
CHAPTER 5: SUMMARY, IMPLICATIONS, AND FUTURE RESEARCH	100
Summary and Conclusions	100
Incidental versus Intentional Word Learning	100
The Influence of Prior Word Knowledge on Metacognitive Monitoring and Calibration.....	101
Processes Reported During Reading.....	103
Text Factors Impacting Reading Comprehension and Metacognitive Monitoring	106
Limitations	107
Implications and Significance.....	109
Research.....	109
Educational Practice.....	111
Future Research	113
APPENDICES	115
Appendix A: Word Knowledge Pretest	115
Appendix B: Narrative Passage	125
Appendix C: Word Knowledge Posttest	135
Appendix D: Think Aloud Protocol.....	143
Appendix E: Practice Passage.....	144
Appendix F: Coding Scheme for the Verbal Protocol Analysis	145
Appendix G: Demographics Sheet.....	148
Appendix H: Word Characteristics from WordNet 2.1	149
Appendix I: Coding Scheme for Main Idea Response Ratings	152
Appendix J: Calibration Scatterplots	153
REFERENCES	155

LIST OF TABLES

Table 1	Descriptive Statistics of Metacognitive Monitoring and Word Knowledge	60
Table 2	Intercorrelations between Metacognitive Monitoring, Word Knowledge, and General Reading Skills	61
Table 3	Summary of Regression Analysis for Metacognitive Monitoring Variables Predicting Change in Word Knowledge	62
Table 4	Context Clues Available in the Passage for Each Target Word	70
Table 5	Means and Standard Deviations for All Variables in the Path Analysis	85
Table 6	Covariances for All Variables in the Path Analysis.....	86
Table 7	Summary of Data Model Fit Indices.....	87
Table 8	Think Aloud Data	95

LIST OF FIGURES

Figure 1	Fukkink's Process Model.....	22
Figure 2	Proposed Model of Word Learning and Calibration of Word Learning.....	77
Figure 3	Model with Standardized Path Coefficients	88
Figure 4	Mean Differences in Main Idea Scores by Passage	97
Figure 5	Mean Differences in JOLs by Passage.....	98

CHAPTER I

INTRODUCTION

Throughout the 1980's a proliferation of studies supported the positive effects of vocabulary instruction on reading comprehension (e.g., McKeown, Beck, Omanson, & Perfetti, 1983; Stahl & Fairbanks, 1986). A report from the National Reading Panel (NRP, 2000) called attention to the paucity of research on effective methods for fostering word learning. The vocabulary and comprehension subgroup of the NRP could not conduct a meta-analysis of vocabulary instruction because there was too much variance in operationalization of vocabulary instruction, methodologies, and implementation. After review of 56 studies, a subgroup of the NRP made tentative recommendations for effective vocabulary instruction, such as providing multiple exposures to new words, and using both direct and indirect instruction.

Two drawbacks to the aforementioned recommendations highlight considerable gaps in our understanding of word learning. First, studies reviewed by the NRP subgroup were not empirical studies of word learning, but studies of vocabulary instruction. Since little is known about word learning processes it is difficult to design instructional programs or interventions to increase gains in word knowledge (Marzano, 2004). Second, NRP's vocabulary and comprehension subgroup did not acknowledge that a substantial portion of word learning happens outside the purview of formal instruction but rather occurs incidentally.

Jenkins, Stein, and Wysocki (1984) define *incidental word learning* as the ability to derive and retain new word information without explicit direction. Because it has been hypothesized that only 10% of new word meanings are learned through direct instruction

(Nagy, Anderson, & Herman, 1987), and because reading accounts for such a large portion of individuals' learning of new words, it is crucial to examine the metacognitive processes that lead readers to recognize the presence of unknown or partially known words in text and to take appropriate cognitive action. Noticing a gap in linguistic knowledge and the concomitant need to locate or infer meaning for unknown words requires metacognitive monitoring and locating or constructing meanings for those words requires regulation of cognition.

According to Flavell (1979) metacognition is "knowledge and cognition about cognitive phenomena" (p. 906). His model of cognitive monitoring categorized metacognitive monitoring as a type of metacognitive experience. Regulation of cognition was classified as an action or strategy. Thus, individuals draw on metacognitive experience as well as actions or strategies when faced with the task of determining word meaning during reading. When discussing the conscious and often effortful processes of metacognition, it is important to consider approaches individuals might take towards the task of meaning-making during reading.

Daalen-Kapteijns, Elshout-Mohr, and de Glopper (2001) described orientations to word learning while reading for comprehension. They described *text oriented* activities as those that a reader engages in to understand the main idea of the text. From this orientation, readers would only derive meaning for unknown words, through strategies such as substitution and checking, if it was necessary to sustain the flow of reading comprehension. On the other hand, *word oriented* activities are those concerned with using context to determine the meaning of unknown words. This would lead readers to a context-specific representation of word meaning that may or may not support future

encounters with the same word in different contexts. Finally, Daalen-Kapteijns et al. described *vocabulary knowledge oriented* activities. These are driven by the goal of increasing vocabulary knowledge and encoding new features to one's mental lexicon. Readers purposefully decontextualize derived aspects of word meaning in order to associate it with what they already know about similar words or morphological parts.

It is the encoding entailed in the last type of orientation that was the basis for Sternberg and Powell's (1983) theory of learning word meaning from context. More recently, researchers (e.g., Bolger, Balass, Landen, & Perfetti, 2008) have posited an instance-based learning approach to deriving word meanings. From this perspective, encounters with words provide information about one or more features, and the context of the encounter is encoded along with those features. This information shapes subsequent encounters with the same word. Over several encounters enough information accrues, and associations become strengthened enough to abstract certain core features that constitute a decontextualized understanding of the word's meaning.

Although the instance-based learning approach describes the mechanisms for learning new word meanings, it does not indicate which processes are necessary for successful versus unsuccessful encoding. As Daalen-Kapteijns et al. (2001) suggest, not every reader will engage in the encoding of word features the same way. The current study suggests calibration as a metacognitive indicator to understand perceived success with word learning. Poorly calibrated learners do not recognize when they have overestimated or underestimated their performance on outcome tasks that demonstrate learning. Studies of knowledge calibration and reading comprehension calibration have found that learners most often overestimate their performance (Glenberg & Epstein,

1985; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstien & Fischhoff, 1977; Yates, Zhu, Ronis, Wang, Shinotsuka, & Toda, 1989).

If individuals believe they have encoded features of word meaning more accurately than they actually did, they may use misleading information to guide subsequent encounters with the same word. They may activate inaccurate prior knowledge, fail to employ appropriate strategies to extract further information or correct inconsistencies, or fail to engage in help-seeking behavior. Calibration is an important consideration related to word learning because well-calibrated learners are presumably aware of what they know and do not know, which allows for more accurate encoding of word meaning features (Zimmerman, 1990).

In short, the basis for change in word knowledge depends on readers' orientation to word learning within context and also their ability to abstract semantic features from word and text level information (Bolger et al., 2008; Daalen-Kapteijns et al., 2001). Little is known about how exactly that process unfolds, and even less is known about the metacognitive processes necessary to calibrate the encoding process. Fukkink (2005) provided one model for word learning that represents a series of metacognitive activities in which learners engage when faced with an unknown or unfamiliar word. First readers make a hypothesis as to a word's meaning, next readers check the fit of the hypothesized meaning, and the check leads to an evaluation to either accept or reject the hypothesized word meaning. Fukkink considers this a recursive process since several word meanings are often generated and tested in succession. According to Fukkink's model, readers would then check the fit of this new concept within the sentence and find that it is an acceptable meaning because it fits all constraints within the sentence. If that fit is not

good then readers must engage in the whole process again until they find a good fit between meaning and context.

One limitation to this model is that it only assumes a text-oriented approach (Daalen-Kapteijns, Elshout-Mohr, & de Glopper, 2001) to determining word meaning while reading; that is, readers only derive word meaning if it is essential to rapidly and fluently understanding the main idea. Another potential limitation is that this model has only been empirically tested with middle school children. However, the components of the model seem appropriate for undergraduates because formal training in word meaning acquisition ceases or significantly decreases around 4th grade (Graves, 2006). Therefore, adults often engage in similar processes to those specified by Fukkink to determine the meaning of an unknown word. This will be evident in the similarities between studies examining adult processes (Durso & Shore, 1991) and school children's processes (McKeown, 1985) to be discussed at length in Chapter Two. Although this model begins to illuminate metacognitive processes related to word learning, alternative models are needed to examine how reading comprehension and metacognitive monitoring influence word learning from text.

Sternberg and Powell (1983) hypothesize six metacomponents in their theory of word learning, one of which they called "monitoring the solution to the problem" (p. 888). They suggest that recognizing success or failure when learning new word meanings relates to the problem of recognizing and correctly utilizing information to infer word meaning. The current study sought to examine person variables (reading skill, prior word knowledge, passage comprehension, and judgments of learning) related to learning word meanings from context and calibration of word learning.

When word learners calibrate (judging their confidence) they ask themselves, “*How well* did I understand the meaning of that word?” This is slightly different than a *judgment of learning*, where word learners would ask themselves, “Did I understand the meaning of that word?” Judgments of learning are global evaluations of whether or not encountered material has been retained and can be recalled at a later time (Dunlosky, Serra, Matvey, & Rawson, 2005). Calibration is a calculated indicator of how *accurate* learners’ judgments were relative to their actual performance on a measure of word learning. Specific word, text, and person factors have been found to influence incidental word learning during reading. For example, Schwanenflugel, Stahl, and McFalls (1997) found that grammatical part of speech and conceptual complexity accounted for differential gains in word knowledge for both unknown and partially known words. Contextual support and text difficulty have been found to influence ability to derive word meaning from surrounding text (Beck, McKeown, & Caslin, 1983; Swanborn & de Glopper, 1999). Thus, when context is supportive enough expert readers are capable of encoding features of word meaning from a single contextual exposure (Bolger, Balass, Landen, & Perfetti, 2008).

Statement of the Problem

The current study addressed several gaps in the literature on word learning and calibration. First, the majority of word learning studies have focused on children (Fukkink, 2005; McKeown, 1985; Nagy, Anderson, & Herman, 1987), even though adults face increasing word learning demands in college courses, advanced technical training, and reading newspapers and popular books (Curtis, 2006). Vocabulary knowledge has long been found to significantly influence reading comprehension

(Anderson & Freebody, 1981; Davis, 1944). Second, many word learning studies have used artificially constructed texts and tasks that do not reflect typically encountered texts, and therefore have limited generalizability to the way adults engage in everyday word learning (Swanborn & de Glopper, 1999). Third, metacognition has not been extensively studied in relation to word learning, although it is increasingly being studied in relation to reading comprehension (Thiede, Dunlosky, Griffin, & Wiley, 2005).

It is important to address these research gaps because adults tend to be poorly calibrated with regard to a range of cognitive tasks (Lichtenstein & Fischhoff, 1977; Lin, Moore, & Zabrocky, 2001). This is a potential roadblock to successful word learning, and therefore successful reading comprehension. Examining the influences on word learning and calibration of word learning has the potential to inform instructional practice. Training has been shown to improve calibration for a specific task (Lichtenstein & Fischhoff, 1980), and therefore could be explicitly taught to improve metacognitive evaluation for struggling word learners. In order to inform instructional practices, more must be known about influences on calibration of word learning after reading.

An alternative to Fukkink's (2005) model has the potential to clarify metacognitive monitoring during and after word learning. Because Fukkink's model is based on deliberate word learning, the current study assigned participants to either an intentional or incidental word learning condition. This was done to provide empirical evidence to allow researchers to determine whether or not it is appropriate to infer parity between learners' approaches to these types of word learning. In-depth consideration of related person, text, and word factors will increase our understanding of how to provide

supportive modeling through teaching, text quality, and approach to vocabulary instruction.

It has been well documented that development in academic domains is predicated on students' ability to acquire a base of conceptual knowledge and that individuals' vocabulary is an effective indicator of that knowledge base (Alexander, Murphy, Woods, Duhon, & Parker, 1997). It has also been well documented that students' metacognitive awareness (i.e., knowledge of self as a learner and thinker) is significantly related to academic development (Veenman, Elshout, & Meijer, 1997). Yet, what is less well understood is the degree to which metacognitive awareness predicts word learning. The purpose of this study was to investigate that relation between metacognition and word learning for competent readers.

It is often difficult to disentangle prior topic knowledge from prior word knowledge in expository texts since most low frequency (target) words are technical terms (Gardner, 2004). Narrative texts, on the other hand, decrease the confounding of prior topic knowledge with prior word knowledge. Low frequency words in narrative texts are typically not technical; instead they are difficult synonyms of already known concepts (Gardner, 2004). Texts in the current study were selected as suitable contexts to study students' change in word knowledge in texts of appropriate difficulty. Once processes and products of word learning are better understood within narrative texts, the same processes can be examined within expository texts, which entail the additional complexity of prior topic knowledge interacting with prior word knowledge.

Purpose of the Study

The current study sought to examine undergraduates' metacognitive processes during word learning, a crucial component of building representations of main ideas, and key concepts from text (Stahl, 1999). There were two main goals for the study. The first goal was to assess empirical evidence relative to a model of the influences of prior word knowledge, reading skill, passage comprehension, and metacognitive monitoring on word learning and calibration of word learning using methods that directly measured word learning and metacognitive evaluation. The second goal of the study was to determine which text factors influenced the ease with which word learners could derive meaning while reading and evaluate their level of performance on a word knowledge test.

Research Questions and Hypotheses

Research Question 1: Does a model of word learning and calibration of word learning that accounts for influences of prior word knowledge, reading skill, passage comprehension, and passage JOLs have better fit for a plain text condition or a bolded keywords condition?

Literature on word learning from text has suggested that incidental and intentional word learning are distinctly different approaches to word learning (Swanborn & deGlopper, 1999). Empirical studies have investigated the differing outcomes of incidental versus intentional word learning (Konopak, Sheard, Longman, Lyman, Slaton, & Atkinson, 1987), and the cognitive and metacognitive processes readers report while intentionally learning words from text (Fukkink, 2005).

While the differences between incidental and intentional word learning have been widely accepted in word learning research, there is no known research that directly tests a

model of word learning with groups randomly assigned to an incidental or intentional word learning condition. It was hypothesized that the proposed model of word learning would fit differently for the intentional word learning condition (key-word bolded) than for the incidental word learning condition (plain text). In particular the influence of JOLs on word learning outcomes and the influence of JOLs on calibration were expected to differ across the two groups.

Research Question 2: Does prior word knowledge relate to judgments of learning and calibration?

The hypothesis for this question was that higher levels of prior word knowledge would be positively related to metacognitive monitoring and evaluation for undergraduate students. Individuals with higher levels of prior word knowledge might be expected to have gained that knowledge through strategic processing, and therefore be more practiced at effective metacognitive monitoring and regulation of cognition (Fukkin & de Gloor, 1998; McKeown, 1985).

Research Question 3: Are specific kinds of processes and strategies that participants report while thinking aloud related to calibration of word learning and changes in word knowledge?

This is an exploratory question, since think aloud data have not been collected to date in empirical research on calibration. Frequency and descriptive data were analyzed after transcripts were coded based on Pressley and Afflerbach's (1995) verbal protocols for metacognition while reading.

Research Question 4: Are there differences in participants' passage main idea scores across narrative passages?

It was hypothesized that increasing text difficulty (as indicated by the Lexile framework) would decrease participants' ability to accurately state the main idea of the passage. This is an important question because if readers have difficulty constructing a situation model for the story they will not be able to pay attention to finer grained details such as individual word meanings within the text (Kintsch & van Dijk, 1978).

Research Question 5: Are there differences in participants' passage judgments of learning across narrative passages?

The hypothesis for this question was that judgments of learning would not decrease with more difficult passages. Several studies have demonstrated that students are not very well calibrated to particular tasks and that overconfidence is the largest source of bias (Glenberg & Epstein, 1987).

Definitions of Terms

Vocabulary is formulaically defined by some linguists as, "a set of W pairs (f, s), where a form f is a string over a finite alphabet, and a sense s is an element from a given set of meanings. Each form with a sense in a language is called a word (Miller, 1995). Individuals have several vocabularies depending on context (oral vs. written), task (receptive vs. expressive), and domain. *Receptive vocabulary* can be demonstrated with a sentence decision task, allowing participants to discriminate between correct and anomalous usage of a word. *Expressive vocabulary* can be demonstrated through the ability to retrieve a word from one's lexicon that appropriately conveys the meaning of a desired concept (Durso & Coggins, 1991). Thus, word knowledge from expressive vocabulary is more explicit and well-formed than word knowledge represented by receptive vocabulary.

Word knowledge is a complex and ill-structured representation that allows one to apply a word “flexibly but accurately in a range of new contexts and situations”

(Anderson & Nagy, 1991; p. 721).

Partial word knowledge is a representation of the meaning of a word that includes both correct and incorrect attributes (Fukkink, 2005).

Incidental word learning is the ability to derive and retain new word information without explicit direction (Nagy, Anderson, & Herman, 1987).

Reading comprehension is "the process of simultaneously extracting and constructing meaning through interaction and involvement with written language" (p. 11). Further, there are three aspects of comprehension: the reader, the text, and the task (RAND Reading Study Group, 2002).

Metacognition refers to individuals’ knowledge about cognition and monitoring of their ongoing or recently completed cognitive processes, such as remembering and comprehension (Flavell, 1979).

Baker and Brown (1984) include Flavell’s metacognitive monitoring in a group of activities called *regulation of cognition*. Other activities include checking, planning, revising, and evaluating.

Calibration is the accuracy with which learners assess their knowledge on a particular cognitive task (Glenberg & Epstein, 1985; Lichtenstein & Fischhoff, 1977).

Judgments of learning are decisions about the likelihood of being able to demonstrate knowledge of recently encountered information on an assessment (Dunlosky, Serra, Matvey, & Rawson, 2005).

Significance of the Study

This study was undertaken with three main goals, each addressing a major gap in the existing literature on word learning from text. First, there are a few studies that examine how word learning unfolds during reading (Bolger et al., 2008; Fukkink, 2005), yet surprisingly little empirical evidence exists for which aspects of reading influence the process of word learning or the product of vocabulary knowledge. The endeavor to create a model of hypothesized influences on word learning drew attention to the overwhelming lack of existing theoretical perspective on how aspects of reading comprehension might influence word learning from text. Therefore the current study established theoretical rationale for a testable model of influential factors in word learning from text.

Second, metacognitive monitoring and calibration of reading comprehension have recently become an area of major investigation (Lin & Zabrucky, 1998; Wiley, Griffin, & Thiede, 2006). There has been a growing recognition that it not only matters if and how students successfully comprehend a given text, but also how well they perceive they achieved success in comprehending the text and how that perception impacts future reading comprehension (Lin, Moore, & Zabrucky, 2001). Since the importance of metacognitive monitoring and calibration have been recognized in relation to reading comprehension, it was deemed empirically interesting to determine the potential implications of these constructs for word learning, especially since word learning during reading would heavily rely on perceptions of learning and success. Therefore, the hypothesized model for the current study was designed with metacognitive monitoring

(i.e., judgments of learning or JOLs) and calibration of word learning (assessed via confidence ratings) as critical components.

Third, there has been a paucity of research on the confluence of features relating to text difficulty that are present in naturally occurring texts (i.e., those written by authors to be read by anyone who will buy the book, not those written by experimenters for a specific study). Thus, different features of text have been manipulated in the study of word learning for various purposes. In an effort to learn about typical word learning, texts for the current study were chosen rather than constructed. Several different features of each text were analyzed to determine text difficulty and passage comprehension and monitoring were compared across passages. Utilizing naturally existing texts and including analysis of text difficulty increase the validity and generalizability of the results.

CHAPTER II

LITERATURE REVIEW

The purpose of this review is to provide a framework for studying calibration of word learning. The first section addresses the word learning literature, with a view towards what is known about word learning processes, how word knowledge changes are typically measured, and existing gaps in the literature that need further examination. The second section addresses one of those major gaps, the need for research on the link between metacognition and word learning. More specifically, this review will consider an aspect of metacognitive evaluation called calibration. This construct is considered from a historical perspective and from the perspective of research paradigms commonly used to investigate self-evaluations of reading comprehension. Finally, the last major section will examine person, text, and word factors that influence word learning, and are presumed to influence calibration of word learning. All the factors reviewed were examined with think aloud data from a subset of participants in the current study.

Word Learning

It is estimated that the average student learns approximately 3,000 new words per year in grades 3 through 12, and has a 40,000 word vocabulary by the time they finish 12th grade (Nagy & Anderson, 1984). Research has shown that only 300 of those 3,000 words per year can be learned through direct instruction in school (Stahl, 1999). Students learn the other 90% of words incidentally from oral and written contexts such as speech, classroom or home environment, school reading, free reading, and television (Nagy & Herman, 1987). Jenkins, Stein, and Wysocki (1984) define *incidental word learning* as the ability to derive and retain new word information without explicit direction (Nagy,

Anderson, & Herman, 1987). Because only 10% of new word meanings are learned through direct instruction, and because reading accounts for such a large portion of students' learning of new words, it is crucial to examine the cognitive and metacognitive processes students go through while determining the meaning of unknown words encountered in text during a typical reading situation.

Numerous studies have examined students' incidental word learning while reading (Konopak et al., 1987; Nagy, Herman, & Anderson, 1985; Nagy et al., 1987; Swanborn & de Glopper, 2002). For example, Konopak et al. (1987) investigated whether or not grade 11 readers would notice and construct meanings for unknown words encountered in content area texts without explicit instruction. Noticing unknown word meanings requires metacognitive monitoring and constructing meanings for those words requires regulation of cognition. Participants were in one of 3 groups. The intentional word learning group received text with highlighted words and was asked to define each target word after reading the passage. The incidental word learning group received the same text without highlighted target words, and the control group received a different text that did not contain the target words. All groups took a pretest and posttest where they indicated their knowledge of each target word (yes or no) and generated a definition for each word.

Findings indicated that the intentional learning group had the greatest gains in word knowledge and confidence in words they thought they knew from pretest to posttest. The incidental word learning group made smaller gains. More importantly however, the incidental word learning group did make gains in word knowledge. Their responses were more general and contained more inaccuracies than those of the

intentional word learning group, but they were able to generate a significantly greater amount of correct target word definitions than the control group (Konopak et al., 1987).

These findings suggest that even a typical reading situation, where difficult words are not highlighted; students learn information about word meanings. Nagy et al. (1987) estimated that children encounter approximately 16,000-24,000 new vocabulary words within a million words of running text. They further calculated that if children learn only 5% of those words, a conservative estimation, they would gain 800-1,200 new word meanings over the course of typical reading.

Nagy et al. (1987) measured incidental word learning by examining the accuracy of generated definitions for target words. Konopak et al. (1987) made a distinction between the specific and accurate definitions written by intentional word learners, and the more general and at least somewhat inaccurate definitions of incidental word learners. Such a difference illustrates that incidental word learning is an incremental process and relates to the idea of partial word knowledge. Durso and Shore (1991) noted that earlier studies of partial word knowledge failed to accurately assess level of word knowledge because researchers did not clearly define and measure partial word knowledge. According to Durso and Shore there are three levels of words: unknown words, frontier words, and known words. Readers are unable to distinguish unknown words from made-up words. Frontier words are words that readers recognize as real because they have been encountered before, but claim to be unable to define. They are called frontier words because they are on the edge of readers' knowledge. Readers can generally place frontier words in the correct general context, even without knowing word meaning. It is possible for readers to define known words and understand their meaning within multiple

contexts. Although these distinctions have been made, little consideration has been given to how metacognitive activities during reading relate to these levels of word knowledge.

Understanding of partial word knowledge has come from research asking students to deliberately derive word meaning from context (Fukkink, 2005; McKeown, 1985). Partial word knowledge is a representation of the meaning of a word that includes both correct and incorrect attributes (Fukkink, 2005). This is one phase of the meaning acquisition process described by McKeown (1985) as the “discovery of a stable meaning for an unfamiliar word that makes sense in, and illuminates the meaning of, the contexts in which the words appear” (p.484). In order to move from partial knowledge of a word to a stable and accurate knowledge of meaning, the word must be encountered in multiple contexts.

A connectionist model can serve to explain how knowledge about a word accumulates through repeated exposures. Within a connectionist model a concept is represented by a node and each node has links to other nodes that are related or have similar properties. This implies that a concept is the whole network of nodes that stem from the initially activated concept node (Collins & Loftus, 1975). During the first exposure to a word, the word’s orthographic information would be linked to the context in which it is encountered. A word would be remembered in the following ways: “It has something to do with...” “I remember seeing it in...” This kind of representation is not generalizable because it is context-bound.

After repeated exposures to a word, information found in multiple contexts strengthens some nodes and weakens others. A word is defined by its nodes or features (Stahl, 1991). To put it another way, new information from each encountered context is

added to existing knowledge and the representation of word meaning one has in memory. Eventually there is a saturation point at which the word can be understood and used appropriately in every context. According to the connectionist model understanding of a word's meaning should grow at a relatively constant rate (Schwanenflugel, Stahl, & McFalls, 1997). More accurately, understanding of a word's meaning should grow at a constant rate depending on context, characteristics of the word, and person features (such as prior knowledge) that will be the focus of the latter part of this review.

A more concrete way to describe the ideas of word meaning acquisition from the connectionist framework can be found in the tasks devised to investigate the Meaning Acquisition Process (MAP). McKeown (1985) designed a study to investigate the differences word learners, at varying ability levels, had in constructing meaning from context. She constructed a 5-step meaning-acquisition task for 30 Grade 5 students. Students were chosen based on their 4th grade Vocabulary subtest scores on the Stanford Achievement Test. The high performing group consisted of 15 students who had a score at grade level equivalent to 4.8 or above. Students in the low performing group had a score at grade level equivalent to 4.1 or below.

Additionally these low-performing students had a score at grade level equivalent to 3.3 or higher on the Reading subtest in order to ensure absence of serious reading problems. First, a sentence containing a pseudoword, the target word, was presented to participants with 6 choices for possible word meaning. Students were asked why they thought each definition choice was appropriate or not appropriate for the target word. Second, two more sentences with the same target word but different contexts were presented with the same 6 choices for word meaning. Students were again asked why

they thought answers were appropriate or not. Third, three sentences similar to the step 1 sentence, but with different details (some giving no additional information) were presented. Students were asked if these sentences gave any more information about the correct word meaning. Fourth, students were asked what they thought was the word meaning. They were given another sentence with “strong” context if they guessed incorrectly. Fifth, six sentences with the target word were presented and students were asked whether the word was used correctly or incorrectly in each sentence.

Although McKeown (1985) did not examine metacognitive activities as part of her study, she asked questions of her participants that reflect its importance to the MAP. At almost every step participants were asked to indicate the appropriateness of their definition choice. In order to comment on the appropriateness of word meaning, students had to engage in some sort of evaluative reasoning. Although metacognitive activities were not a part of McKeown’s analyses and findings, she drew conclusions about individual differences in word learning that emphasize a need for instruction in both word learning strategies and metacognitive activities such as monitoring.

McKeown (1985) found that students in the low-performing group did not grasp the relation between the word and its context. They had difficulty using the clues provided by the sentences to construct and refine a meaning for the word, and struggled to use the word in new contexts. This suggests that the meaning acquisition process and the connectionist model of word learning it represents are only as effective as the skills the reader brings to the task of deriving word meaning. If readers struggle to connect unknown words to context in which they appear, they will not benefit as much from

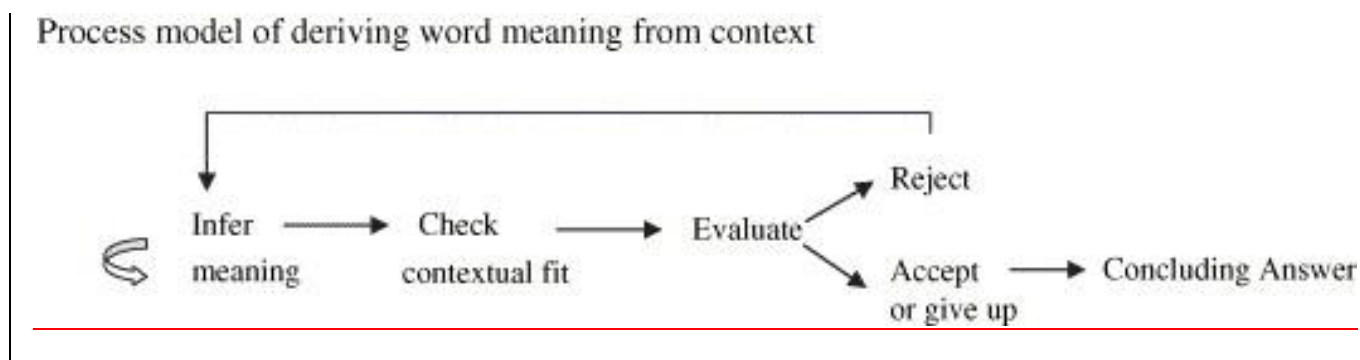
incidental word learning. Remembering that 90% of words are learned incidentally, it is easy to see what a disadvantage low-performing readers would have.

Learning from contextual clues only happens if there is some loose understanding of a word already. For example, readers might use morphological clues to determine some aspects of meaning and constrain the underlying concept to a certain category (i.e., object, action, descriptor, etc.). The combination of that knowledge, gained from a lifetime of exposure to oral and print words, with knowledge gleaned from contextual clues available in the text, leads to word learning if the reader chooses to pay attention to the unknown word and utilizes appropriate strategies to extract information and derive meaning (Sternberg & Powell, 1983).

Target words are indicated in an intentional learning task, and directions usually indicate that participants should either derive meaning for those words, or that they will be tested on those word meanings at a later time. Sometimes participants are also instructed in how to derive meaning from context. This focuses attention, and subsequently metacognitive monitoring. Incidental word learning on the other hand depends on individuals' choice to pay attention to unknown words. If the choice is not made to acknowledge lack of word knowledge (e.g., comprehension does not suffer from lack of knowledge for a few words) then processes of word learning will not unfold.

A Model for Word Learning

Once readers have searched for clues and determined constraints upon word meaning from context they are ready to infer word meaning based on all collected information. Fukkink (2005) offers a model that begins at the step of inferring word meaning (see Figure 1).

Figure 1. *Fukkink's process model*

First readers make a hypothesis as to a word's meaning, next readers check the fit of the hypothesized meaning, and the check leads to an evaluation to either accept or reject the hypothesized word meaning. Fukkink considers this a recursive process because often several word meanings are generated and tested in succession. For example, in the sentence "Sirius threw aside the bread he had just lifted to his mouth and instead picked up the flask of pumpkin juice and drained it" (Rowling, 2000, p. 529), if *flask* were an unknown word to 5th grade readers of this text, they would have access to a word level clue, the fact that *flask* is used as a noun. Further, context within the sentence indicates that one can drink from a *flask*, that it holds juice and that one can pick it up. From these clues a reader might hypothesize that a *flask* is a container one drinks from, like a cup or a bottle. Readers would then check the fit of this new concept within the sentence and find that it is an acceptable meaning because it fits all constraints within the sentence.

Fukkink (2005) provided support for the model by collecting data from 40 students in grades 2, 4, and 6. Students were given identical texts manipulated to control

for word and text effects. The texts were designed to be 100 word narrative texts of high frequency words for each grade level. Within those texts were 12 low frequency target words, half abstract and half concrete. No explicit clues were provided for the words in the surrounding text. In effect, there were no synonyms, antonyms, or description clues, and all 12 target words were indicated in bold typeface, making this an intentional word learning task as opposed to an incidental word learning task. Students were instructed to talk aloud about what they were doing to determine the meaning of the bolded words as they were reading. Definitions provided by the students were scored based on the percentage of correct attributes given within the definition. False attributes, or characteristics not related to a word, were also scored dichotomously.

Fukkink found that think-aloud data from the students reflected all stages of the model. Students did not necessarily follow the model in the sequence designated, but rather flexibly used the activities as each situation demanded. Although Fukkink investigated deliberate word learning, the model should work similarly for incidental word learning. The major difference between deliberate and incidental word learning is that when a reader is not forced to learn specific definitions they are unlikely to enter into the processes described by McKeown (1985) or Fukkink (2005) at all unless the word is essential for comprehension of the passage.

McKeown (1985) and Fukkink (2005) also investigated specific phases of Pressley and Afflerbach's (1995) four phases for word-related activities. In the first phase, readers decide whether or not to spend effort determining a word meaning. If readers decide to spend effort figuring out a word meaning, they engage in phase 2, where they pay greater attention to the word and its context. In other words, they search

for clues about word meaning. Sternberg and Powell (1983) named this process selective encoding, or finding relevant information from the context in order to determine word meaning. This phase was further delineated by McKeown's (1985) Meaning Acquisition Process. After searching for clues, readers infer word meaning with the use of context. They make a guess about what meaning best fits the sentence. Phase 3 is referred to as selective combination in Sternberg and Powell's (1983) processes of knowledge acquisition. Readers must use a combination of clues presented in context in order to make an appropriate guess about an unknown word's meaning. Readers also compare clues from context to prior knowledge about the topic or situation. Sternberg and Powell refer to this as selective comparison. Finally, in phase 4, readers evaluate the fit of that generated meaning. Phases 3 and 4 are part of Fukkink's (2005) model of word learning.

The evaluation process within Fukkink's (2005) model relates to the judgments learners make about new word meanings and their level of confidence about future encounters with those words in different contexts. This is likely the most crucial aspect of Fukkink's model, when one thinks of it in terms of success with incidental word learning, since readers are typically not directly tested on their knowledge of unknown words encountered while reading. They rely instead on judging what is right and wrong with their best guess of a definition.

The strong relation between reading comprehension and vocabulary knowledge has been found repeatedly since the inception of research on reading processes (Stahl, 1999). Adult readers continue vocabulary growth, as reading is an essential to participation in our literate society (Alexander, 2005). Studies suggest that educated native English speakers acquire approximately one or two words per day, and have a

receptive vocabulary of around 17,000 words (Goulden, Nation, & Read, 1990). Vocabulary knowledge is a necessary, but not sufficient, requirement for reading comprehension. Having a large vocabulary does not ensure reading comprehension, but it positively affects readers' attempts to more fully comprehend a text. Another necessary, but not sufficient, requirement for successful reading comprehension is an awareness of cognitive processes used to monitor progress toward the goal of understanding a piece of text, check outcomes, and redirect failed efforts. Thinking about cognitive processes is called *metacognition*, and this phenomenon also has been widely studied in relation to reading comprehension since its more focused conceptualization and subsequent operationalization from the 1970's to present (Baker & Brown, 1984; Flavell, 1979; Lin, Moore, & Zabucky, 2001).

Significantly less attention has been given to how metacognition influences the word learning that leads to vocabulary knowledge. Models of word learning exist that incorporate metacognition (such as Fukkink, 2005) into a process model of how readers learn the meaning of unknown words. The next step in understanding metacognition's affect on word learning is to examine specific metacognitive activities and their function within a process model, or working understanding of how a reader determines meaning for an unknown word. This review considers aspects of the calibration of reading paradigm as they might be applied to studies of word learning calibration, or in other words the goodness of fit between readers' prediction of the accuracy of word learning and their actual performance on a word learning accuracy test. Conceived in this manner, calibration is a specific part of the evaluation piece of metacognition.

Metacognition

Metacognition refers to individuals' knowledge about cognition and monitoring of their ongoing or recently completed cognitive processes, such as remembering and comprehension (Flavell, 1979). According to Baker and Brown (1984), knowledge about cognition is stable general information one has about a task such as reading. For example, a reader might understand that a passage full of unfamiliar words and concepts will be more difficult to comprehend than a passage with familiar words and concepts. Baker and Brown include Flavell's (1979) metacognitive monitoring in a group of activities called *regulation of cognition*.

If readers check their guess of a possible word meaning for its fit within the sentence, they are engaging in a regulatory process. Regulation of cognition is not stable like knowledge of cognition. It is situation-specific or condition-specific. For example, a reader may have knowledge of a particular strategy, but fail to employ it due to high cognitive load or lack of interest (Baker & Brown, 1984; Lin & Zabrucky, 1998). Therefore, knowledge of cognition does not automatically lead to regulation of cognition in a particular situation.

In their review of metacognition and reading comprehension, Lin and Zabrucky (1998) further divided regulation of cognition into two components: evaluation of comprehension and regulation of comprehension. Evaluation of comprehension happens when, "a reader becomes aware of his or her comprehension difficulties" (p. 346). Regulation of comprehension is when a reader uses a strategy to overcome comprehension difficulty. Both components seem readily transferable to word learning

and have already been examined in several studies (Konopak, Sheard, Longman, Lyman, Slaton, & Atkinson, 1987; McKeown, 1985).

A commonly investigated aspect of evaluation is one's judgments of learning (JOL). A JOL occurs when readers realize they do not understand what they have just read, or when learners make a decision about whether they are ready to be tested on something within the text, such as target words (Dunlosky, Serra, Matvey, & Rawson, 2005). JOL has some overlap with calibration. Readers must make a JOL in order to calibrate their learning. However, calibration is more specifically how accurately a JOL represents reality.

The accessibility hypothesis of judgment accuracy describes why individuals tend to be poorly calibrated in their JOLs. Dunlosky, Rawson, and Middleton (2005) define the accessibility hypothesis as:

People's judgments are inferential in nature, and in particular, such inferences are based on the total amount of information accessed immediately prior to making each judgment. An important assumption is that people often do not (or cannot) evaluate whether the information accessed is correct or incorrect, so only the quantity and not the quality of the accessed information is expected to influence metacognitive judgments. (p. 552-3)

They found that when participants incorrectly represented information during recall, they tended to overestimate JOLs. This finding lends support to both the accessibility hypothesis and previous findings of individuals' tendency to be poorly calibrated (e.g., Glenberg & Epstein, 1985).

Think-aloud protocols were collected from high-skilled and low-skilled undergraduate readers in the Netherlands (Daalens-Kapteijns & Elshout-Mohr, 1981). The protocols were coded for use of model-building in deriving meaning of a pseudoword. They found that high-skilled readers analyzed features of word meaning in each context and made an abstraction from their accumulated knowledge from all contexts. Low-skilled readers, however, extracted a meaning feature from the first context, and proceeded to fit information from each of the following contexts into that initial model of word meaning. Interestingly, they did not significantly differ in their demonstration of word meaning through written definitions.

The processes of word meaning derivation uncovered by Daalen-Kapteijns and Elshout-Mohr (1981) indicate different patterns of strategy use, as manifested in the different kinds of models used to derive meaning from multiple contexts. They also indicate differential metacognitive monitoring and calibration of word learning. One approach was to integrate information from features presented across contexts, while another approach was to fit all information into the feature(s) derived from the first context encountered. This finding resonates with the previously described study McKeown conducted with 5th grade high-skilled and low-skilled readers. She also cited an interference effect across multiple contexts for the less-skilled readers, and a lack of metacognitive monitoring to notice and repair inconsistencies.

Sternberg and Powell (1983) described reading skill level as one of several variables that mediate the process of learning word meaning from context. Among these mediating variables is the variability of contexts surrounding the target word, the importance of the target word to the overall meaning of the text, the supportiveness of the

context, and the usefulness of prior knowledge. Many of these variables are directly addressed in the current study. More emphasis needs to be placed on person variables such as metacognitive monitoring. Perhaps high-skilled and low-skilled readers differ due to their metacognitive monitoring and not their level of reading skill. This is especially plausible since definitions generated by high-skilled and low-skilled readers do not appear to differ and low-skilled readers did not demonstrate difficulty in comprehending the sentences presented as context (Daalens-Kapteijns & Elshout-Mohr, 1981).

Research Paradigms

There are several ways that researchers have captured readers' judgments about their best guesses to an unknown word's meaning. Konopak et al. (1987) simply asked participants to indicate their perceived knowledge of particular words by answering yes or no after the pretest and again after the posttest. Maki, Shields, Wheeler, and Zacchilli (2005) studied reader predictions, performance, and confidence more extensively than Konopak et al. Maki et al. used several kinds of perception scales to investigate metacomprehension accuracy in adults. Participants read 6 passages and after each passage were asked to report how much text they believed they successfully comprehended. They indicated this with a scale of percentages that ranged from 0-100% in 20% increments. Additionally they predicted how many (out of 6) test questions they would answer correctly for each passage. Finally, participants answered 36 questions (6 for each passage) and indicated how many (out of 6) they believed they answered correctly for each passage.

Maki et al. (2005) found that low-performing readers, as identified through z-score conversion of scores from the verbal portion of the SAT, were overconfident in their performance predictions on difficult texts. They predicted they would answer more questions correctly than they actually did answer correctly. High-performing readers, however, were found to accurately predict their test performance, but they were underconfident in their posttest reports of performance. While studies such as Maki et al.'s have begun to uncover patterns in readers' judgments about learning and comprehension, rating scales cannot capture processes involved in making those judgments.

In order to examine individual differences in judgments of learning, some researchers have made use of think-aloud protocols. Fukkink (2005) employed think-aloud protocols to determine whether or not participants followed his model of deriving word meaning from context. Participants were instructed to determine the meaning of a difficult word within the text and to talk aloud while doing so in order to let the researcher know how the participant accomplished this task and to which features of context and memory the participant paid attention.

After transcribing each participant session, Fukkink divided each protocol, or verbalization, into utterances in order to distinguish between semantically distinct guesses or answers. There were 5 categories within the coding scheme that matched the process model: infer meaning, check contextual fit, evaluate, reject or accept, and concluding answer. Each utterance was coded as one of those categories. There were 360 protocols total from all participants, most with multiple utterances to be coded. Notice that think-aloud protocol methodology is labor-intensive. Data must be collected

and recorded during the participant session. Then it must be transcribed and coded before it can be analyzed in any way. Once these steps are accomplished, however, a rich data set exists that allows for unique analyses. For example, Fukkink performed a sequential analysis of his think-aloud protocols in order to determine whether or not patterns were evident that supported the sequence of his model.

The major drawback to this methodology is that it cannot be applied to studies of incidental word learning because the procedure disrupts typical reading activities. Participants were asked to deliberately notice a difficult word and derive its meaning. This is quite different than most reading situations where no direction is given as to what to pay attention to and what to learn from a given text. For this reason, the less-intrusive, albeit less informative, rating scales are preferred for incidental word learning studies.

Calibration

Similar research paradigms are used to investigate a particular metacognitive skill called *calibration*. Recall that metacognition has 2 components, knowledge and regulation; that regulation has 2 components, evaluation and regulation; and that metacognitive evaluation is called calibration. Calibration is an individual skill influenced by cognitive, motivational, and contextual factors. Readers make evaluations based on their goals for reading, their perceptions of task difficulty, and their comfort with the text. Calibration is the accuracy with which students assess their knowledge of a cognitive task. Calibration has been often measured through the administration of confidence scales and ratings of understanding in conjunction with a reading comprehension task (Glenberg & Epstein, 1985; Maki & Berry, 1984; Weaver, 1990). Calibration is quite different from JOLs because JOLs are participants' self-reported

judgments, while calibration is a calculation of the relation between participants' self-report judgments and their actual performance and this relation indicates accuracy of judgments.

Lichtenstein and Fischhoff (1977) considered calibration to be one of 3 measures of *probability assessments*, or statements of confidence in one's knowledge. They also measured *over/underconfidence* and *resolution*, or the ability to discriminate varying degrees of uncertainty within a set of items. Resolution was considered independent of calibration, and a formula that included knowledge, calibration, and resolution produced the *Brier score*. This could be calculated to obtain a measure of performance adequacy on a given set of questions.

In order to determine the influence of expertise and item difficulty Lichtenstein and Fischhoff (1977) conducted a study of undergraduate and graduate students within the domain of psychology. They found that participants who knew more about the given questions demonstrated better calibration and resolution. This was only true up to a certain point however, because they also found that participants who answered more than 80% of the questions correctly were more poorly calibrated than the group that only got 70% correct. Another finding suggested that regardless of expertise, participants showed higher calibration on easier items, and that calibration worsened with increasing item difficulty. Finally, they found that people tend to be overconfident regardless of expertise, but tend to be somewhat underconfident for the easiest items.

Lichtenstein and Fischhoff (1980) also conducted a study to determine the effect of training participants to improve their calibration. Training sessions consisted of discussion of what calibration was and how it was important in decision-making and

performance; feedback in the form of calibration curves, frequency of probabilities assigned, and proportion of correct answers; and discussion at the end of each session where participants attempted to explain what they were learning and how they applied that to calibration of knowledge tasks. Findings showed an improvement in calibration after just one training session. Thus, they concluded that performance feedback for individuals who were initially poorly calibrated increases the accuracy of their probability judgments.

Another potential influence on calibration was considered almost a decade later – culture (Yates, Zhu, Ronis, Wang, Shinotsuka, & Toda, 1989). They found that Japanese and American participants were better calibrated, although still overconfident, than Chinese participants. They also found that Chinese participants demonstrated better resolution than American and Japanese participants. These results were especially interesting because there were no significant differences in proportion of correct answers across cultural groups.

This early body of work on calibration suggests that it is a rational judgment, meant to reflect the state of one's knowledge at the time of answering a question, or completing a task. None of the studies described explicitly mentioned calibration as part of metacognition. However, a study by Koriat, Lichtenstein, and Fischhoff (1980), exploring reasons why adults tend to be poorly calibrated in the direction of being overconfident, was included in the metacognitive monitoring section of a book entitled *Metacognition: Core Readings* (1992). Since then, researchers have included calibration in models of metacognitive processes as the indicator by which individuals monitor their

planning and goal setting. This is an entirely different conceptualization than that of earlier calibration research.

One feature of calibration is that it reflects an internal judgment of correctness than can be measured against external criteria of agreed upon correctness. Calibrating confidence in knowledge to demonstrated knowledge is quite a different situation than calibrating plans and goals to an individually determined standard. Furthermore, those individual standards can be assumed to vary due to epistemic cognition, motivation, and prior knowledge for tasks within particular domains.

Readers make evaluations about whether or not a generated word meaning fits based on judgments of their efforts and the extent to which those efforts achieve the desired goal of comprehension. Notice that this two-fold evaluation is based on perceptions. To make an accurate evaluation, there must be a strong correspondence between one's perceived efforts at word learning and actual success using the word in new contexts. This is extremely important to regulatory activities of word learning because regulatory activities are invalid if readers' metacognitive awareness inaccurately represents achievement and word learning activities.

Researchers who study vocabulary acquisition recommend repeatedly exposing students to target vocabulary, and guiding students in making connections between new words and prior knowledge (National Reading Panel, 2000). For exposure and connection-making to be effective, however, students must accurately *calibrate* their word learning: that is, they must determine whether or not they have successfully learned word meaning. In the literature, calibration is determined by the correspondence between readers' perceived success learning a word and their actual success defining that word or

using it in a new context. Glenberg and Epstein (1985) noted that, “a well-calibrated individual correctly assesses his state of knowledge, knowing when he knows, and knowing when he does not know. In contrast, the self-assessments of knowledge of the poorly calibrated person are uncorrelated with actual states of knowledge” (p. 703). Unless students are aware they do not truly understand the meaning of a word, repeated exposures to that word will not foster learning, and may lead to misrepresentation of the word and its underlying concepts.

Lin, Moore, and Zabucky (2001) conducted a study on comprehension calibration (Lin et al., 2001). Sixty undergraduate students read 12 expository paragraphs with true-false inference questions after each paragraph. They also completed comprehension confidence scales before reading and confidence scales to measure predictions of performance after the paragraphs but before the inference questions. Lin et al. found that calibration of comprehension and calibration of performance was generally of low frequency, although calibration of performance was slightly more frequent than calibration of comprehension. Results from the study (Lin et al., 2001) indicate that college students poorly calibrate their learning efforts.

Findings that college students show poor calibration of comprehension support earlier findings from Glenberg and Epstein (1985). In experiment 1 of their study, 85 college students were presented with 15 paragraphs. After reading each paragraph, they responded to a confidence rating scale (1 = very low to 6 = very high) in their ability to use information from the text to draw inferences about its topic. They then completed a true/false inference verification task, where they responded to the correctness of an inference related to the main idea of the paragraph. The purpose of this step was to

ensure that participants were judging the same material for both the confidence rating and the main idea inference. Additionally, students were grouped by when they responded to a confidence scale rating their performance on the inference test. The immediate condition group rated their confidence after each text, and the delayed condition group rated their confidence on all the texts at the end of the session.

Glenberg and Epstein (1985) measured calibration of comprehension in three ways. First they calculated a point biserial correlation between posttest confidence scores and actual performance on the inference questions. They reported very low correlations, suggesting that overall calibration of performance (called calibration of comprehension in the article) was poor across subjects. Second, they calculated the confidence-judgment accuracy quotient (CAQ) in order to determine the relation between the pretest confidence rating and actual performance on the inference questions. The mean CAQ scores were extremely low, suggesting poor calibration of comprehension. Finally, Glenberg and Epstein calculated calibration curves to illustrate the proportion of correct inference answers with the 6 levels of confidence possible on the posttest confidence scale. The slopes for both the immediate and delayed conditions appear shallow, indicating poor calibration. All three calculations point to low levels of calibration of comprehension, a finding supported by Lin et al. (2001) and Winne and Jamieson-Noel (2002).

Another important set of findings from Lin et al.'s (2001) study was the criteria students use to calibrate their learning. Participants completed four pretest ratings of understanding of text, confidence in ability to answer questions on read material (JOL), easiness of texts, and interestingness of texts. Multiple measures of calibration of

comprehension reflect considerations readers must make about task, personal, and text factors when they engage in calibration of comprehension. Lin et al. found that the four pretest measures were related to each other and that students who were highly calibrated on one of the measures tended to be highly calibrated on all the other pretest measures as well. This is a notable finding given that studies have demonstrated that text difficulty (Maki et al., 2005) and level of interest in particular texts (Alexander & Jetton, 1996) impact reading comprehension. Lin et al.'s (2001) study offers preliminary evidence to support that these factors also correspond to calibration of comprehension.

Calibration is conceptually and operationally distinct not just from JOL's, but also from self-efficacy. The differentiation between self-efficacy and calibration depends on *when* they are measured. Self-efficacy is task-specific, and Pajares (2002) states that self-efficacy impacts one's choices. For example, one will choose tasks in which one feels competent. Therefore, it is often measured *before* a particular task is administered (Bong & Skaalvik, 2003) and can be a predictive judgment. In this way, self-efficacy judgments are based on one's initial beliefs about a task. Calibration is defined as the relationship between confidence and actual performance (Nietfeld, Cao, & Osborne, 2005). For this type of judgment to be made, it must occur *after* a task. Hence, calibration judgments are based on reflective abstraction of the task itself.

Future studies of calibration of reading processes, whether reading comprehension or word learning, should consider multiple factors of influence as Lin et al. (2001) did in their study on calibration of reading comprehension and performance. Critical factors will vary by type of reading process examined. However, in general it seems that what is important for the reading process being examined is also important in the consideration

of calibration of that process. When investigating calibration of word learning and performance person, text, and word factors should be considered.

Word Learning Variables Studied in Relation to Calibration

Calibration research has not yet been applied to word learning, although Konopak et al. (1987) included a confidence measure in their study. Konopak et al. reported that participants in the incidental word learning group reported they knew more words than their scored definitions reflected. Overconfidence, or what Glenberg and Epstein (1985) refer to as illusion of knowing, has consequences for future encounters with partially known words. A positively biased evaluation of word learning may cause a reader to decrease self-monitoring of word learning tactics (Pintrich & Zusho, 2002). However, Durso and Shore (1990) reported that participants were accurate in identifying correct and incorrect usage of words they reported as unknown to them. They explain this phenomenon as support for readers' implicit knowledge of general word constraints. Even if general knowledge about words is not dependent on accurate metacomprehension, or more specifically JOL's, an awareness of those factors is necessary to make the kind of sentence decisions required in Shore and Durso's (1990) task. Distinction between person and word factors is an important consideration in studying calibration of word learning.

Choosing factors to examine in relation to calibration of word learning is guided by factors commonly examined in both word learning and calibration of comprehension studies. Sternberg (1987) delineated processes of knowledge acquisition, or person factors, types of contextual clues, or text factors, and a variety of mediating variables related to both text and person. Schwanenflugel, Stahl, and McFalls (1997) studied the

contributions of text factors and word factors to word learning. Stahl (1991) summarized the importance of person, text, and word factors in vocabulary learning. Given this body of previous research, and the review of calibration of comprehension research conducted by Lin and Zabrucky (1998), numerous factors seem central to studies of calibration of word learning. Person variables to be considered are prior knowledge, purpose for reading, and interest of the information or task to the reader. Text variables to be considered are text type or structure, text difficulty, and quality of context. Word variables to be considered are importance of the word to the main idea, conceptual complexity, and morphological clues.

Person factors.

Prior knowledge. Readers' prior knowledge is brought to bear when searching for clues within text (Waern, 1988). Readers use knowledge about words and the different ways that knowledge can be used in order to derive word meaning. Prior knowledge about words includes idiomatic usage of words, such as understanding that "muddy the waters" does not refer to wet dirt in a river; recognition that words are polysemous or have multiple meanings, such as tomato plant versus nuclear plant; and realizing that words are often interrelated, meaning knowledge of one word is not independent from knowledge of other words (Graves, 2000; Nagy & Scott, 2000). For example, horse, stirrup, and saddle have meanings that relate to each other, and activation of the concept of stirrup may also activate the concept of horse.

Another kind of prior knowledge is what Shore and Durso (1990) refer to as implicit knowledge of word meaning. Prior knowledge about the interrelated nature of words, multiple meanings of words, and idioms are all explicit knowledge about word

meaning. Implicit knowledge includes information about general constraints of a word rather than specific information about meaning. For example, a reader might not recognize the word *rapacious*, but they would not assume that it refers to a profession because they have some general idea of the constraints of the word. While this section addresses only prior word knowledge for a task, the section on word factors describes the kinds of characteristics inherent to words that readers may or may not have knowledge of when they determine unknown word meanings.

Prior knowledge and calibration of reading comprehension have been studied by Glenberg, Sanocki, and Epstein (1987). They specifically examined the affect of domain-familiarity on readers' accuracy of comprehension calibration. In experiment 3, 88 college students read 15 paragraphs about a variety of topics (e.g., blood sugar and black holes). Depending on group assignment, students then completed either a familiarity rating for specific statements from the text, or recalled certain information from the texts prompted by statements, or did neither. All groups completed confidence ratings on their ability to answer test questions about those specific topic statements and then completed an inference verification test.

Because of the way these data were collected, Glenberg et al. (1987) only examined averages at the group level and did not draw conclusions about individual differences in calibration of comprehension. For each paragraph, they calculated an average familiarity rating, confidence rating, and performance score. They found that domain familiarity, as measured by ratings of familiarity with topic statements, was correlated with confidence, $r = .66$. This evidence was reported to support the claim that confidence is based on domain familiarity.

Nietfeld and Schraw (2002) investigated the effect of prior knowledge and strategy training on calibration of mathematical problem-solving. Prior knowledge, they believed, provides a standard against which to compare calibration judgments and also improves performance. Strategy training, according to Nietfeld and Schraw, enhances calibration by freeing cognitive resources for metacognitive activities that would otherwise be used for problem-solving.

In Experiment 1, Nietfeld and Schraw (2002) tested three hypotheses about how prior knowledge influences calibration. The debilitating hypothesis suggests that prior knowledge is negatively related to calibration. As individuals gain competence within a given domain they become overconfident and calibrate their performance based on self-efficacy within a domain, rather than adjusting for the specific problem. Therefore, high levels of prior knowledge should positively influence performance, but negatively influence calibration. This hypothesis echoes Glenberg and Epstein's (1987) findings.

The no-impact hypothesis suggests that prior knowledge is not related to calibration. It also predicts that prior knowledge will increase performance. Finally, the facilitative hypothesis suggests that prior knowledge improves both performance and calibration. Schraw and Nietfeld (2002) explain the rationale for this hypothesis in terms of expertise. Experts have vast prior knowledge in a given domain, and therefore have a conceptual basis to make their performance evaluations and enough available cognitive resources to allocate to calibration.

To test these hypotheses, Nietfeld and Schraw (2002) divided 93 undergraduates into groups based on their prior knowledge of statistics. All participants completed a general ability test for mathematics, and a 24-item multiple-choice probability problem-

solving test. After each question on the test participants rated how confident they were in correctly answering the question on a 100-point continuous scale. An accuracy score was calculated following the calibration paradigm. It measured the difference between each participant's confidence score and performance score (1= correct, 0 = wrong) for each item. Nietfeld and Schraw (2002) found that the high prior knowledge group scored significantly higher on the probability test and had significantly higher accuracy scores than the other two groups, suggesting they were better calibrated. There were no differences between the low- and mid-prior knowledge groups.

In Experiment 2, Nietfeld and Schraw (2002) shifted their focus to novice statistics students. Specifically they wanted to know what effect strategy training would have on calibration. Strategy training was hypothesized to positively influence calibration for three reasons. First, strategy training brings attention to the importance of monitoring and evaluating performance. Second, once learners use strategies they use fewer cognitive resources for problem solving and are able to devote those resources to metacognitive activities. Third, having strategies when faced with challenging problems increases self-efficacy, which may lead to increased motivation to calibrate problem solving.

Nietfeld and Schraw (2002) administered a general math skills test, the 24-item probability test from the first experiment, and a mathematics self-efficacy questionnaire to 58 undergraduates. In a second session, one group participated in a two hour intervention on strategies for solving probability problems. The control group participated in a two hour strategy training session that was unrelated to mathematics. In

a third session, all participants completed a parallel form of the probability test and the same mathematics self-efficacy questionnaire.

Findings revealed that the groups did not differ in performance on the probability test. Participants in the intervention group reported significantly higher confidence than did participants in the control group. Additionally, participants in the intervention group were significantly better calibrated after training than they were before training. Self-efficacy did not differ across groups, but it was positively related to performance, confidence, and calibration. Self-efficacy was also found to be highly related across sessions.

Nietfeld and Schraw's (2002) findings support their hypothesis that prior knowledge and strategy training increase performance and calibration of mathematical problem solving. Although Glenberg and Epstein (1987) found contradictory evidence about the relation between prior knowledge and calibration of reading comprehension, they studied the variables averaged across groups and not at the individual level as Nietfeld and Schraw did. Recall that several studies have considered the effect of prior knowledge (Durso & Shore, 1990; Konopak et al., 1987; Schwanenflugel et al., 1997) on incremental word learning, and found that some prior knowledge does increase vocabulary test performance. However, prior knowledge has not been examined in relation to calibration of word learning. Nietfeld and Schraw make several important arguments for prior knowledge's facilitative role in calibration in the context of a complex task such as reading. If word learners already have some conceptual knowledge of a word they have a standard to use in evaluative judgments of progress and performance. Also, the more knowledge readers have about the text's general topic and

the semantic features of words, the more automatic a process, such as Fukkink's (2005) model of deriving word meaning from context becomes. As these cognitive processes become more efficient, learners have resources to allocate to monitoring and evaluation such as calibration.

Purpose and interest. Factors closely related to prior knowledge are purpose and interest. Swanborn and de Glopper (2002) examined 6th-graders' purpose, when starting to read a passage and its effect on word learning. Three types of purposes were assigned: free reading, reading for text comprehension, and reading to gain topic knowledge. A fourth group served as the control group and had no assigned reading purpose. Swanborn and de Glopper hypothesized that the free reading group would show the lowest growth in word knowledge because their attention need not be directed at any specific aspects of text or vocabulary to understand the general meaning. They hypothesized that the text comprehension group would show the highest gains because those readers would in fact pay close attention to text and word factors.

Although Swanborn and de Glopper (2002) discuss growth and gains, they did not administer a pretest definition task of any kind. They measured general comprehension and administered the definition task after the assigned goal reading task. Definition accuracy, used to measure incidental word learning, was assumed to reflect new knowledge gained from the texts because the words were assumed to be wholly unknown. This assumption was based on teachers' ratings of the likelihood of each word being unknown to their students.

Despite this serious methodological drawback, the researchers found group differences. The reading for topic knowledge group defined the most words correctly on

the posttest, followed by reading for comprehension, and finally the free reading group generated the least number of correct definitions (Swanborn & de Glopper, 2002). These findings support the notion that incidental word learning varies in quality depending on one's purpose for reading.

Related to purpose and attention is *reading engagement*. Schraw and Bruning (1999) defined reading engagement as “the degree to which readers generate critical and personal responses to text” (p. 282) and considered how beliefs about readers' role influence reading engagement. Two kinds of systems of beliefs were outlined. The *transmission model* assumes that meaning is transmitted from the page to readers' memory, in other words the reader passively receives meaning. Meaning is independent of the reader according to the beliefs held in the transmission model. Conversely, the *transaction model* assumes that meaning resides in readers' minds and must be actively constructed from text. Schraw and Bruning (1999) note that adult readers tend to hold both kinds of beliefs simultaneously, thus influences from both models impact reader engagement. Additionally, implicit models guide readers' goals, strategy selection, and judgments of learning.

Purpose for reading and beliefs about readers' role in gaining information from text would seem to explain individual differences in choosing whether or not to focus attention and cognitive resources on deriving word meaning from context, the starting point for previously outlined process models (Fukkink, 2005; McKeown, 1985; Pressley & Afflerbach, 1996). However, these factors are insufficient without some level of interest in the topic, or in the task. Situational interest, such as curiosity about a particular topic, has been linked to increased attention, persistence, text recall, and

learning according to Hidi's (2001) review of interest and reading. Neither topic interest, nor interest in language or vocabulary, has been studied in relation to word learning or calibration of word learning. Since beliefs, purpose, and interest directly influence motivation to read they also indirectly affect decisions to engage in word learning while reading.

Text factors. Individual differences are not the only factors that account for variation in word learning performance and calibration. Text and word factors determine the amount of information that can be obtained about an unknown word. The ease with which that information is abstracted from text can also be attributed to features such text difficulty and conceptual complexity. The following sections analyze these features, starting with the broad category of genre, and ending with a discussion of specific categories of conceptual complexity.

Narrative vs. expository text. Kintsch and van Dijk (1978) suggest that if a word is part of an idea that is high in the text structure, that is important to the passage meaning, the passage will contain more information about that idea and therefore that word. This is called the text processing model and is related to the notion that finding the meaning of an unknown word is secondary to constructing meaning for the text as a whole when reading (Fukkink, 2005). Expository texts tend to include low-frequency words that are crucial to text meaning as they are often topic-specific technical terms. Narrative texts, on the other hand, tend to include low-frequency words that enrich deep comprehension, but are not crucial to understand the main idea of the text.

Nagy, Anderson, and Herman (1987) found that strength of contextual support was related to word learning from context in expository texts, but was unrelated to word

learning from narrative texts. It is tempting to assume this finding is the result of stronger contextual support in expository texts, which often include definitions and synonyms as clues for low-frequency words, and weaker contextual support in narrative texts. However, Nagy et al., did not find differences in strength of contextual support between expository and narrative texts.

More in-depth analysis of the differences between narrative and expository texts was reported by Gardner (2004), who studied typical 5th-grade texts (28 expository and 28 narrative). She found that genre effects types of words to which students are exposed, number of encounters with specific words, and amount of prior knowledge necessary for word learning. Narrative texts had more high-frequency words and fewer word types, and thus were deemed more facilitative to incidental word learning because they presented fewer lexical demands. Expository texts, however, had more low-frequency words vital to both general and domain-specific vocabulary, and had more repetitions of those words. While lexical demands were higher for these types of texts, the potential for word learning was much higher in expository texts than in narratives.

Gardner's (2004) findings are especially illuminating given Nagy et al.'s claim that strength of context and word learning are similar across genres. If this is the case, and the potential for vocabulary acquisition is highest for expository texts it would be advisable to encourage free reading of nonfiction, newspaper articles, and other themed materials as well as high quality fiction. The only drawback to free reading of expository texts is the higher demand for prior knowledge in order to decrease lexical demands of texts dense with specific or technical vocabulary. Such specific vocabulary is one of the considerations of formulas designed to measure text difficulty.

Text difficulty. Readability formulas have been used by both researchers and educators to approximate reading level for texts. Standardized testing has placed heavy emphasis on students' reading level – whether they read at or below grade level is a chief concern of the *No Child Left Behind* legislation. Grade level reading is often determined by matching students' calculated reading level to age-related norms of reading level. Those age-related norms are matched to readability classifications of texts in most current reading curricula.

Chall and Dale (1995) reviewed the construct of readability and found that the strongest predictors of text difficulty are sentence length and word difficulty. Word difficulty has referred to either low-frequency words, or word length (e.g. syllables). Researchers use readability measures to control the difficulty of the texts presented to participants. Word learning is optimized when a text is somewhat challenging, meaning just above individuals' reading level (Nagy, Herman, & Anderson, 1985; Stahl, 1999). If text is too challenging, readers struggle to meet comprehension demands, and thus do not have enough cognitive resources for word learning from context or metacognitive activities such as calibration. In such an overwhelming situation calibration is expected to be quite poor (Glenberg & Epstein, 1985; Lin, Moore, & Zabrucky, 2001). If text is too easy, readers are unlikely to encounter any unknown words.

Context. Another source of clues is degree of contextual support within a particular text. Degree of contextual support depends on the type of context and the distance of the contextual clues from the unknown word. Context can be classified as directive, generally directive, neutral, or misdirective. Directive context provides strong clues, perhaps even a definition for the unknown word. Few typical texts provide

directive context. Generally directive context provide the reader with some information about possible meanings for words, whereas neutral context do not. Misdirective context provides clues that lead the reader to incorrectly guess the meaning of an unknown word (Beck, McKeown, & McCaslin, 1983). Beck et al. found that more directive contextual support helped adult readers derive word meaning from context. In addition to context type, the strength of relation between the unknown word and main idea is generally a determinant in the amount of information context provides about the unknown word (Kintsch & van Dijk, 1978).

One might assume that undergraduates would be as likely to use a dictionary or glossary as they would to derive word meaning from context. Such a method seems to be more direct and less effortful. However, several studies have directly compared word knowledge gained with the help of definitions to word knowledge gained through context. Findings have shown that definitions were not helpful, or much less helpful than context in acquiring word knowledge (Fischer, 1994; Bolger, Balass, Landen, & Perfetti, 2008). Nagy and Scott (2000) suggest several reasons for this pattern of findings. First, definitions do not provide any information about acceptable usage in context. This supports Bolger et al.'s (2008) instance-based learning approach, where meaning features are first encoded with their context so as to increase associations with future encounters in context and increase speed of activation. Additionally, students often look for a synonym within definitions, thus ignoring the whole definition and only encoding one salient feature of the word's meaning.

Word factors. Fukkink (2005) suggests that readers use 3 kinds of information in their search for clues to the meaning of an unknown word encountered in text: clues

from the word, clues from the context, and clues from prior knowledge. Clues from words come from several sources, including morphological analysis, grammatical category of the word, word concreteness, and conceptual complexity. Morphological analysis is the consideration of word parts such as roots and affixes (Baumann, Edwards, Boland, Olejnik, and Kameenui, 2003). For instance, a reader may know the meanings of the word *fashion* and the suffix *-able* and is therefore able to determine meaning for the more complex *fashionable* when encountered in text. Grammatical category refers to whether a word is a noun, verb, adjective, etc. Grammatical categories of unknown words can usually be determined by the syntactic structure of the sentence and the use of the word in the sentence (Brown, 1957). Word concreteness is the degree to which the concept a word represents is concrete, such as *table*, or abstract, such as *love*. Concrete words have been found to be easier for children to learn (Schwanenflugel, 1991).

Finally, conceptual complexity is the degree to which an unknown word is related to known concepts. Durkin (1990) makes a distinction between simple synonyms, where the unknown word is related to a known concept, and conceptually challenging words, where the unknown word is a new label that refers to an unfamiliar concept. For example, the word *frightened* might be unknown to a reader, but it is very similar to the concept of *scared*. If a reader recognizes that the new word is a verb and that the surrounding context supports the notion that it is similar to the known concept of *scared* they can assume that *frightened* is a synonym and understand the meaning of the sentence.

Schwanenflugel, Stahl, and McFalls (1997) found that word factors were more important than text factors in development of vocabulary knowledge. They first gave 43

4th graders a vocabulary checklist to determine their familiarity with specific words. After a one-week break they tested story comprehension by giving students a 6th grade level text to read and asking them to write a summary of the passage each day for two days. Students were then given a multiple choice definition test with choices that reflected: the correct definition, a domain-relevant partial definition, and 2 incorrect choices. Regression analyses were run with the following predictor variables: word concreteness, grammatical part of speech, number of times a word appeared in text, level of contextual support, and level of text importance. They found for unknown words that none of the text or word factors were significant influences on word learning. For partially known words however, word concreteness and grammatical part of speech were significant influences on word learning. As discussed above, concreteness and grammatical part of speech are word factors; thus supporting Schwanenflugel et al.'s claim that word factors are more important than text factors in vocabulary learning.

These properties are part of a broader classification of word difficulty that has been examined in several ways (Graves, 1984; Jenkins & Dixon, 1983). Jenkins and Dixon (1983) define four conditions for unknown words:

Condition 1: The unknown word (e.g. *formidable*) is a more complex synonym of a simpler word (*difficult*), and the reader knows the concept indicated by the simpler word.

Condition 2: The unknown word (e.g. *subterfuge*) is a more complex synonym of a simpler word (*artifice*), but the reader does not know the concept indicated by the simpler word.

Condition 3: The unknown word (*gill*) is not a synonym for a simpler word, but the reader is familiar with the concept (e.g. what a fish uses to breath).

Condition 4: The unknown word (*okra*) is not a synonym for a simpler word, and the reader is not familiar with the concept.

Graves (1984) classifies unknown words from a different perspective:

Type 1 Words: Words that are part of a reader's oral vocabulary that they are unable to read.

Type 2 Words: New meanings for words a reader already knows and recognizes with other meanings

Type 3 Words: Words that are not part of a reader's oral or reading vocabulary, but for which they can build a concept.

Type 4 Words: Words that are not part of a reader's oral or reading vocabulary, and for which they cannot easily build a concept.

Both of these classification systems consider word difficulty as a function of a reader's familiarity with the concept referred to by an unknown word. Jenkins and Dixon (1983) take the perspective of word difficulty through possible associations with other words, whereas Graves (1984) describes word difficulty in terms of connections to a reader's vocabulary knowledge. For the purposes of research stimuli, the Jenkins and Dixon system allows researchers to rate word difficulty with the use of a thesaurus and a word frequency corpus. The Graves system on the other hand considers word difficulty based on individual vocabulary knowledge. Thus it can only be used to rate target words after participants have been tested to determine whether or not the words are part of their vocabulary.

When choosing or designing stimuli for incidental word learning experiments, word difficulty should reflect what is typically encountered in texts. Nagy, Anderson, and Herman (1987) argue that studies often examine context use with words that do not reflect typical unknown word encounters. Cloze tasks, where the target word is replaced by a blank, and replacing target words with nonsense, or pseudowords, both present problems for studying typical encounters with unknown words in text. The cloze task leaves the reader without clues from orthography (spelling) or morphology (recognizable word parts). Pseudowords may have familiar synonyms. In both cases readers can replace the blank or nonsense word with a known synonym. However, within text a reader often encounters words that cannot be associated with known synonyms and require new conceptual knowledge (Nagy et al., 1987).

Another word factor consideration is how to capture partial word learning. Word learning studies typically use multiple choice word tests (e.g., McKeown, 1985; Schwanenflugel, Stahl, & McFalls, 1997), definition generation tests (e.g., Jenkins, Stein, & Wysocki, 1984; Swanborn & deGlopper, 2002), or both (e.g., Baumann et al, 2003). Multiple choice word tests

Perhaps more important than the use of multiple choice or definition generation task is the method used to score those tasks. Some researchers studying incidental word learning have chosen to score word knowledge measures as dichotomously right or wrong (Nagy et al., 1987; Stahl, 1989), but the majority of studies give partial credit for answers on word knowledge tasks in order to accommodate investigation of the incremental nature of word learning. Swanborn and de Glopper (1999) conducted a

meta-analysis of incidental word learning studies and found that assessments sensitive to partial word knowledge yield findings of higher amounts of incidental word learning.

Research Gaps

The purpose of this review has been to summarize and synthesize research on readers' attempts to calibrate their efforts and a specific kind of effort during reading – word learning. First, metacognitive activities and their importance to word learning research were discussed. Several seminal studies (e.g., Konopak, 1987; McKeown, 1985) measured types of evaluations readers made about word learning and word knowledge. These studies were focused on the cognitive processes of word learning, however, and the metacognitive piece apparent in their design and measures was not thoroughly analyzed or discussed. Future word learning studies must include metacognitive activities more centrally when examining the steps of word learning that produce differing levels of word knowledge, such as those proposed by Durso & Shore (1991).

Fukking's (2005) process model of word learning was introduced in the latter part of the section on word learning. A major piece of the model is evaluation (see Figure 1.). However, evaluation as described by the model is a cognitive mechanism for choosing the best alternative from several guesses for an unknown word's meaning. It is inferred that readers have a certain amount of confidence in the correctness of their guess. That confidence can be more precisely studied through the calibration paradigm. Metacognitive activities drive Fukking's process model from start to finish. Readers must decide to pay attention to particular details from context, monitor their construction of word meaning, and evaluate their progress in gaining knowledge about the unknown

word. Since Fukkink's study directed readers to determine meaning for unknown words, he did not consider the choice a reader must make to enter into the word learning process in the first place. Again, monitoring and evaluation help a reader decide whether or not determining word meaning is crucial to overall comprehension, or various other reader goals. If the word is deemed unnecessary to the overall goal, effort will not be spent engaging in the word learning process.

Next, the purpose of using the calibration paradigm to study a specific metacognitive activity related to word learning was described. A definition generation test given after a reading task is enough to determine whether or not readers successfully engaged in word learning if a pretest was given to determine prior word knowledge. The difference between word knowledge from pretest to posttest is an indication of word learning. Measuring readers' efforts to calibrate their word learning allow researchers to tap into metacognitive evaluation without disrupting incidental word learning. This has been done by asking readers to report their confidence or predicted accuracy on a test once they have finished reading a passage (e.g. Konopak, et al., 1987). Often participants have been asked to rate their confidence in their answers on the reading comprehension test as well (e.g. Glenberg & Epstein, 1985).

The calibration paradigm has only been applied to a limited number of word learning studies. More importantly, very little is understood about what factors contribute to calibration of word learning. Once more is known about what facilitates calibration of word learning, the question of how to improve generalized vocabulary instruction with more focus on metacognitive skills for incidental word learning can be examined.

CHAPTER III

METHOD AND DATA ANALYSES

Pilot Study

Purpose

Before collecting data for the current study, I conducted a pilot study with a comparable sample of students. The purpose for conducting the pilot study was threefold. First, I wished to determine the amount of time students required to complete the measures for each of the two sessions. In order to limit participant fatigue it is crucial to limit session time to 30 minutes or less. Second, I wanted to determine the appropriateness of the measures and, more specifically, whether adjustments were necessary to directions and individual items. Third, I wanted to create a reliable coding system for the word-knowledge pretest and posttest measures.

Participants

Ninety-six undergraduates participated in the pilot study, but data were only analyzed from 60 of the participants. There were several reasons for removing participants from the data analysis. First, a large number of participants completed the first session, but were absent from class, or could not complete the second session. Second, several participants failed to complete a whole section or measure. Third, a few participants were removed because they indicated that they were non-native English speakers on their demographics form.

The students were enrolled in either a human development class, or an education class at a large, public university in the mid-Atlantic region of the United States. Students were primarily juniors (65%) and had an average age of 21.1 years. Eighteen

male and 42 female students participated, and were predominantly Caucasian (58.3%). Similar sample characteristics were expected for the dissertation study.

Measures and Procedure

The measures administered to undergraduates in the pilot study were similar to measures used in the current study. A brief description of measurement changes based on data and feedback from the pilot study will be provided here.

Participants completed the reading comprehension and vocabulary subscales from the Woodcock-Johnson III Diagnostic Reading Battery and the word-knowledge pretest (Appendix A) during the first session. The W-J III DRB reading comprehension subscale involved a series of cloze tasks, where students had to fill in the blank with the appropriate word for each sentence. This kind of task is more commonly associated with vocabulary than reading comprehension, although the two are correlated (National Reading Panel, 2000). For this reason, it was determined that the Nelson-Denny Reading Test would provide a better measure of standardized reading comprehension in the dissertation study.

The W-J III DRB vocabulary subscales were a series of association tasks where a word was presented and participants were asked to provide a synonym for the synonyms subscale, an antonym for the antonyms subscale, and the appropriate word for the analogies subscale. The most common feedback received at the end of the first session was that those tasks were too difficult. Additionally, participants took too much time in answering the subscales, which is a concern for participant fatigue.

During the second session one week later, participants read six counterbalanced passages and completed global judgments of learning scales after each passage

(Appendix B). Next, they completed self-efficacy scales, asking "If you were asked to define *boding*, how confident would you be in the accuracy of your response?" for each of the thirty target words. Finally, participants completed the word-knowledge posttest with a confidence scale for each item.

Again, this session exceeded the intended time of 30 minutes. Participants took up to 45 minutes to complete all the measures. For this reason, it was determined that four passages rather than six would be sufficient for the dissertation study. The passages that will be deleted for the dissertation study also address an item-level problem that arose. Specifically, some of the target words showed restricted range when responses were coded. For example, *filigreed* was relatively unknown by most participants both at pretest and at posttest. Therefore, it was deleted from the word-knowledge tests in the proposed study along with associated texts. On the pretest, eight of the target word responses had a restricted range of 0-1. On the posttest, only three target words met the same criteria. These words were discarded for the dissertation study since they did not show any variability of responses. This results in a word-knowledge pretest of only 40 words, a passage reading section of four passages, and a word-knowledge posttest of 20 words for the dissertation study. In this way, issues of time and item reliability were simultaneously addressed.

Six of the target words had limited distribution of scores, however, this was largely due to the number of non-responses that were coded as zero. Forcing completion of all words should increase range of responses for these items. Variance on the pretest ranged from .17 to 1.43. On the posttest, variance ranged from .26 to 1.91. Better distribution and variability of scores were expected once instructions were added

regarding: a) modeling how to write a definition and b) requiring participants to complete all words. The latter was made easier in the dissertation study by implementing computer administration, as participants were not permitted to continue without typing some sort of answer.

Cloze tasks were also administered to ensure that target words chosen for the pretest and posttest were not too difficult. The passages were presented with blanks for the target words and students were asked to fill in the word they thought should go in the blank. From this information it was possible to determine that the target words represented a range of difficulty. Some of the words were easy for students to supply (or their near synonyms), while other words were difficult.

Participants gave a wide range of responses to the passage judgment of learning questions during the pilot testing. The mean ratings on the scale ranged from 25.83% to 93.83% across participants (Table 1). This suggests that overall, participants were fairly confident that they had comprehended the meaning of the passages. It would be helpful to include a brief comprehension question after each passage to determine whether or not actual reading

Table 1

Descriptive Statistics of Metacognitive Monitoring and Word Knowledge

	Person			Word		
	Min.	Max.	Mean (SD)	Min.	Max.	Mean (SD)
Judgments of Learning	25.83	93.83	70.96 (15.89)			
Confidence Ratings	0.80	72.53	28.73 (18.66)	7.33	55.24	27.69 (12.60)
Calibration ₁	0.31	56.97	14.63 (13.89)	0.02	0.62	0.37 (0.15)
Word Knowledge Pretest	4	35	16.87(8.16)			
Word Knowledge Posttest	0	54	15.67(10.23)			
WJ III DRB	73	100	85.28(6.19)			

Note. Calibration was calculated as absolute accuracy for questions across people and relative accuracy across words ($N = 60$).

comprehension matches perceived reading comprehension. This could be done by asking participants to summarize what they have just read in one or two sentences.

The relation between JOLs and confidence ratings was stronger for more difficult passages ($r = .34, p < .01$ to $r = .43, p < .01$) than for the easier passages ($r = .27, p < .05$ to $r = .34, p < .01$). However, there was no significant relation between mean calibration and JOLs ($r = .81, p > .05$), suggesting that they are distinct constructs (Table 2).

Table 2

Intercorrelations between Metacognitive Monitoring, Word Knowledge, and General Reading Skills

	1	2	3	4	5	6	7
1. Comp	—						
2. Vocab	.49**	—					
3. JOL	.10	.30*	—				
4. PCR	.16	.38**	.46**	—			
5. Bias	-.12	.17	.28*	.81**	—		
6. AbsA	-.05	.22	.37**	.81**	.81**	—	
7. WKC	.18	.24	.24	.14	-.28*	-.04	—

Note. Comp = Nelson-Denny Comprehension; Vocab = Nelson-Denny Vocabulary; JOL = Judgment of Learning; PCR = Posttest Confidence Rating; Bias = Bias; AbsA = Absolute Accuracy (calibration); WKC = Word Knowledge Change ($N = 60$).

* $p < .05$, ** $p < .01$

To further test this assumption a regression analysis was run to determine how each variable contributed to change in word knowledge. Results showed that both JOLs and bias were significant predictors of gains in word knowledge (Table 3).

Table 3

Summary of Regression Analysis for Metacognitive Monitoring Variables Predicting Change in Word Knowledge (N = 58)

Variable	<i>B</i>	<i>SE B</i>	β
Judgment of learning	.01	.01	.29*
Bias	-.01	.01	-.36**

Note. $R^2 = .15$, $p < .05$.

* $p < .05$, ** $p < .01$

After analysis, it was determined that the self-efficacy scales used in the pilot study were probably not measuring self-efficacy for word learning. They were presented at the wrong time in the procedure and presenting them during the first session is problematic because it betrays the purpose of the study, causing participants to focus on words rather than reading more naturally. It is also problematic that the question for the self-efficacy scale was worded so similarly to the confidence scale question. This may have primed participants for the calibration task and altered the results of the pilot study. I decided to omit the self-efficacy scales for the dissertation study based on these theoretical and measurement issues.

The pilot study also revealed minor issues with the way directions were presented for the confidence scales on the word-knowledge posttest. Data collected from several participants in the pilot study had to be dropped from analysis due to failure to complete the calibration scales. This could be attributed to the placement of the scale directions on the word-knowledge posttest. If participants were not reading the directions carefully

they may have missed the confidence question. Another plausible reason participants skipped the scales could have been fatigue. The two sessions of the study were lengthier than intended, and participants' generated definitions tended to decrease in both quantity and quality towards the end of the knowledge posttest. The dissertation study could avoid this problem by utilizing shorter versions of all measures described.

For the dissertation study, additional directions were added to the passage section of the second session measures to include an opportunity for participants to demonstrate their comprehension of the passage. Since they were asked to make a judgment of learning (JOL) on their understanding of the passage, this is a concrete task that complements the JOL's.

The last conclusion drawn from the pilot study is that the coding schema for the word-knowledge tests appeared reliable. The researcher coded all responses to target words, and two additional raters each scored one-third of the target word responses. A calculation of Cohen's Kappa index of interrater reliability revealed 85% interrater reliability. This calculation of interrater reliability is corrected for chance agreements, and is therefore a conservative estimate (Cohen, 1968). The same coding schema will be used for the dissertation study.

Method

Participants

Three hundred and nine ($N = 309$) undergraduate students from a public university in the mid-Atlantic region participated in the study. As in the pilot study, undergraduates were the focus of this study for various theoretical reasons. For one, there is a paucity of research on this population. In addition, they rely heavily on their

word learning skills in order to learn new concepts across domains. They were recruited with permission of instructors during class time and were offered extra credit for their participation or an alternate activity at the discretion of their instructors. The sample consisted of both male ($n = 67$) and female ($n = 247$) students of various majors.

Participants ranged in age from 18 to 47 years ($M = 20.47$, $SD = 2.57$) and represented the diverse student body from the larger university (60.8% Caucasian, 15% Black, 15% Asian, 5.4% Hispanic, 1.3% Native American, and 2.5% identified themselves as another ethnicity such as Jewish or Pacific Islander). Participants reported having completed a range of nine to 186 credits ($M = 74.6$, $SD = 25.79$) and also reported having a GPA ranging from two to four ($M = 3.26$, $SD = .42$).

All participants were asked to indicate whether or not English was their native language. Non-native English speakers were not excluded from participating in the study, however their data were only included in the sample of 309 participants if they scored within one standard deviation of the sample mean score on the Nelson-Denny Reading Test. This was done to try and control for potential difficulties in completing heavily linguistic tasks that might arise from not having enough experience with reading and writing in English. Twenty-five participants of the sample of 309 indicated that they were non-native English speakers.

Measures

Nelson-Denny Reading Test. Reading comprehension and vocabulary were measured with parallel forms of The Nelson-Denny Reading Test (NDRT). Since the NDRT is a valid and reliable nationally normed measure of component reading skills it was used to indicate general reading skill of participants. The purpose of the NDRT, as

stated by the authors (Brown, Fishco, & Hanna, 1993), is to assess vocabulary, reading comprehension, and reading rate. However reading rate was not recorded for the purpose of the current study.

There are two parallel forms for the NDRT that have been normed specifically on four-year college undergraduates. Each form includes an 80-item vocabulary test and a seven passage, 38 question reading comprehension test. Participants' scores on the vocabulary portion of the NDRT ranged from 33 to 80 ($M = 63.78$, $SD = 8.04$), $\alpha = .87$, and their scores on the reading comprehension portion of the NDRT ranged from six to 38 ($M = 32.90$, $SD = 3.94$), $\alpha = .70$. The grade equivalent for these mean scores is 16.6, indicating that individuals halfway through their 16th year of schooling would on average perform similarly on the NDRT (Brown, Fishco, & Hanna, 1993). Fifty percent of the sample completed Form G, 48% of the sample completed Form H, and form information was missing for 2% of the sample.

Word-knowledge pretest. To assess participants' prior knowledge of the target words, I created a knowledge pretest for the study ($\alpha = .83$). The word-knowledge pretest consists of a list of 40 words (see Appendix A). Twenty-five target words were chosen from the passages administered in session two. These were words that occur less than ten times per 5 million words of running text, as determined by Carroll, Davies, and Richman's (1971) *The American Heritage Word Frequency Book*. Example target words are *banter* and *dilapidation*. Nine more words were chosen from text surrounding the passages, but not appearing in the passages presented to participants. Example filler words are *admonish* and *capacious*. The purpose of the filler words was to prevent participants from focusing on target words that they would see again in session two.

Finally, six pseudowords from an existing study (Schwanenflugel, Stahl, & McFalls, 1997) were added to the word-knowledge pretest. Pseudowords follow English language rules for orthography, but have no meaning. Example pseudowords are *devernal* and *edarthic*.

Target words were chosen with consideration for part of speech and conceptual complexity. Previous work has found that it is easier to derive meaning for nouns than for other parts of speech (Brown, 1957). For this reason, the current study sought to balance the number of nouns and non-nouns to analyze differences in both word learning and calibration based on part of speech. Conceptual complexity was described by Jenkins and Dixon (1983) in terms of synonyms and familiarity. Words can be categorized as more complex synonyms of either known or unknown concepts indicated by simpler words. It is also possible that words do not have a synonym, and may indicate a concept that is either known or unknown to learners. *Tarn* is an example of the latter situation (see Appendix H).

The directions given to participants directed them to "Write a definition or short description for every word that you can on the list. Please make your definitions as clear as possible so that I know that you understand the meaning of the word. I am not interested in the number of words that you know, so just do your best." After participants completed this first phase of the pretest, they were instructed to "Go through the list again and place a check mark beside any word that you left blank if you have seen it before or if it is familiar to you, even if you are not quite sure what it means." The purpose for this set of instructions was to gain information about partial word knowledge participants may have for target words. The pseudowords forced participants to discriminate between

words they may have previously encountered, and therefore know some semantic feature of, and words that they have never encountered and do not have meanings.

Responses to the word-knowledge pretest were scored on a scale of 0 to 4 with half points. All three scorers were trained to use WordNet 2.1 (Fellbaum, 1998) and a thesaurus (www.thesaurus.com). WordNet 2.1 was deemed a better source to judge semantic relatedness than the dictionary because it generates synsets, or sets of synonyms for a given word. These synonyms are organized by sense (i.e., type of meaning, like the dictionary) and they are also organized hierarchically. This means that it is possible to determine superordinate (hypernym) and subordinate directionality. The dictionary simply provides key features. Semantic overlap in WordNet 2.1 depends on a shared superordinate plus shared semantic features. For example, if the target word is *robin*, its superordinate is *bird*. In this case *thrush* and *bluebird* are coordinate terms because they share the superordinate *bird*. Note that *emu* does not appear on the coordinate terms list because it is not a small songbird. While the superordinate is important, coordinate terms are also determined with consideration of overlapping semantic features. Using this system removes a great deal of scorer subjectivity entailed in using the linear relations of the dictionary and thesaurus method. The thesaurus was still utilized in order to cross-check synonyms or coordinate terms as being either first-order or second-order synonyms.

A score of four was given for responses that capture multiple features of meaning. Multiple meanings required that the coordinate terms or synonyms be found under different subheadings, not simply any two synonyms listed for the word. In the case of words without multiple meanings, four was given if the response had sufficient

specificity. For example, participants who defined *steed* as a horse, or a male horse were not scored a four, but participants who defined a steed as a noble horse, or a spirited horse received a score of four. A score of 3.5 was given to responses deemed in between a three and four.

A score of three was given to direct definitions or first-order synonyms. When the response given used a phrase or sentence that was equivalent to a first-order synonym, but did not contain any of the actual words listed in the WordNet 2.1 definitions, coordinate terms, or thesaurus synonyms, the response was scored a 2.5 out of four. A score of two was given for indirect synonyms, which were found by either searching synonyms ordered by frequency for one of the coordinate terms in WordNet 2.1, or by searching for synonyms of first-order synonyms from the thesaurus. Responses indicating or including antonyms to the target word were scored a 1.5 because while their meaning was completely wrong, it was also strongly semantically associated with the target word. A score of one was given for some correct feature of word meaning.

A half point was given to any target words with a check mark on the pretest or an indication that the word was familiar without a definition response on the posttest. Pretest and posttest half points were used as a familiarity check, but were changed to zero for analyses as the check mark response is not equivalent to the open-ended response given at all other levels of the scoring scheme. Half points were retained if the score was given for a very weakly related and somewhat correct definitional feature. A zero was given for incorrect answers. The deviation between rater responses was small ($SD = .43$) and was acceptable using a more conservative measure of interrater agreement ($\kappa = .46$).

The relation of the word-knowledge pretest to the word-knowledge posttest ($r = .821, p < .001$) and the vocabulary portion of the NDRT ($r = .695, p < .001$) provide evidence of concurrent validity for the word-knowledge pretest. It was especially important that this researcher-made instrument relate to the standardized vocabulary measure as they should both tap general vocabulary knowledge to some extent.

Narrative passages. Four narrative passages approximately 250 words in length were used to present the target words in typically encountered context (Appendix B). The passages were taken from two sources, *The Tales of Edgar Allan Poe* (2004) and *The Complete Works of Washington Irving* (1978). These books were selected as sources because narratives were written by famous American male authors of roughly the same period. Based on text readability, a typically-performing college sophomore could comprehend about 75% of text written by Washington Irving with ease, and 95% of the text written by Edgar Allan Poe. Text readability, often referred to as text difficulty, was determined by the Lexile Framework for Reading (2004). Lexiles are based on semantic difficulty (word frequency) and syntactic complexity (sentence length). Directions on the screen with the first passage directed participants to, “Please read the following passages carefully to determine the overall meaning and the meaning of the bolded words.” For the plain text condition the directions were, “Please read the following passages carefully to determine the overall meaning.” Each passage was followed by instructions to summarize the main idea of the passage in a couple sentences.

The main idea statements provided data on participants’ level of reading comprehension for each narrative passage. The coding scheme ($\kappa = .72$) was as follows: main ideas were coded as 1 if they had enough of the features included in the main idea

agreed upon by the raters and did not include any major incorrect features; main ideas were coded as .5 if there were some correct features and perhaps an incorrect feature, or a somewhat vague description; and a main ideas were coded as 0 if they were completely incorrect (Appendix I).

Existing narrative texts were utilized in the current study in order to increase generalizability. Empirical work on word learning has chiefly used artificially constructed texts and tasks in order to create experimental manipulations (Durso & Shore, 1991; Fukkink, 2005; McKeown, 1985). By manipulating text, researchers change the characteristics of target words, contextual support, and text difficulty. Changing these factors does not simulate word learning opportunities in typically encountered texts. Therefore, it is important to study word learning in a manner which reflects a task undergraduates are likely to encounter over the course of typical reading.

Table 4

Context Clues Available in the Passage for Each Target Word

Passage	Word	Clue(s)	Helpfulness
Night Ride	Clove	Causal	Directive
	Endeavored	Equivalence - antonymy	Neutral
	Lag	Causal	Directive
		Equivalence	Directive
	Stave	Class membership	Neutral
	Steed	Equivalence	Directive
Functional descriptive		Directive	

Passage	Word	Clue(s)	Helpfulness
Walk Home	Blundering	Stative descriptive	Neutral
	Con	Enablement	Directive
		Value	Neutral
	Roost	Functional	Directive
	Rustling	Causal	
		Spatial	Directive taken together
Wended	Spatial	Neutral	
Old House	Dilapidation	Equivalence	Directive
	Fissure	Spatial	Neutral
		Stative descriptive	Neutral
	Masonry	Spatial	Directive, distant
		Stative descriptive	Neutral
	Scrutinizing	Value	Somewhat directive
	Tarn	Class membership	Neutral
School Rivals	Animosity	Class membership	Directive
	Anomalous	Overall passage	Depends on comprehension
	Banter	Functional	Misdirective
		Value	Directive
	Motley	Equivalence	Directive
		Stative descriptive	Directive
Sentiment	Equivalence	Directive	
	Class membership	Somewhat directive	

Note. The coding scheme for the clues was described by Sternberg and Powell (1983). The helpfulness of each clue was categorized using Beck, McKeown, and McCaslin's (1983) scheme.

Judgment of learning scales. After reading each passage, participants were presented with two judgment of learning scales (Appendix B). The first question asked, "How confident are you in your understanding of the passage's overall meaning?" ($\alpha = .808$ for all four passage level questions). The second question asked, "How confident are you in your understanding of the individual word meanings from the passage?" ($\alpha = .808$ for all four word level questions, $\alpha = .893$ for all eight JOL questions). Participants responded by marking a slash on a 100-mm line with 0% at one end and 100% at the other end. Because the 100-mm line is presented on the computer, it is actually a 100-pixel line. Participants could see the number value when they clicked to place the mark along the scale. If they wished to change their rating they were able to do so. The value in using continuous rating scales rather than categorical scales has been demonstrated in the literature (Albaum, Best, & Hawkins, 1981; Schraw, Potenza, & Nebelsick-Gullet, 1993) and was deemed the best way to capture individual differences in self-report of judgments of learning.

Since the JOL scales were so highly correlated ($r = .79, p < .001$), participants' passage and word JOLs were averaged together to represent a single JOL score. This score will be referred to as a passage JOL because it represents judgments about individual passages.

Word-knowledge posttest. The word-knowledge posttest is similar to the word-knowledge pretest, with a shorter format and slightly different directions (Appendix C). Specifically, the posttest consisted of only the 25 target words, without the filler words

and pseudowords ($\alpha = .793$). Participants were instructed to, "Write a definition or short description for each word. Please make your definitions as clear as possible so that I know you understand the meaning of the word. If you are unsure of a word's meaning, write your best guess." Responses were scored on the same 0 to 4 scale as the pretest. The word-knowledge posttest was also found to be highly related to the vocabulary portion of the NRDT ($r = .747, p = .00$).

Confidence scales. A confidence scale followed each word on the posttest (Appendix C) for a total of 25 confidence scales ($\alpha = .92$). The question asked, "How confident are you in the accuracy of your response?" Participants generated a definition, or best guess description for each target word, and then evaluated the accuracy of their response from 0% to 100% on the confidence scale. Confidence ratings will be correlated with actual performance (posttest definition scores) to determine participants' calibration of word learning. Calibration was calculated using the formula for Spearman's rho to correlate scored word-knowledge posttest responses with reported confidence ratings, $M_\rho = 0.36$, range: -0.61 to 0.84.

Think-aloud protocols. In addition to the variety of methods that were utilized for the current study, think-aloud data were collected from a subsection of the sample. Participants indicated on their consent forms if they were willing to be recorded while they verbalized their thought processes during the experimental tasks administered in the second session. Twelve participants were chosen to complete the second session and their verbalizations were digitally recorded.

According to Pressley and Afflerbach (1995), there are several aspects of word learning, as encompassed in reading comprehension, individuals monitor while engaged

in reading. It is these types of monitoring processes that participants were expected to indicate as they thought aloud during while reading the passages. They might verbally indicate *perceptions* during reading and judging their learning or confidence, such as difficulty of the text, linguistic characteristics of the text, whether the text is ambiguous, and the relationship between background knowledge (or lack thereof) and the text. They might verbally indicate *meaningful processing of text*, such as behaviors or strategies for processing challenging words, effectiveness of those behaviors or strategies in determining meaning, progress in determining meaning, ease or difficulty of determining meaning, and success or failure in determining meaning. Word learners might verbally indicate *problems* during reading and monitoring, such as unfamiliar terms in the text, failure to learn the meaning of a word, lack of prior knowledge hindering word learning, and inconsistency between expectations for meaning and information encountered in the text. They might verbally indicate *activation of processes to meet word, text, and task demands*, such as decisions to skip parts of the text, decisions to pay more or less attention to portions of the text, decision to focus on certain words at the expense of others due to high processing demands, and decision to reread portions of the text (Pressley & Afflerbach, 1995, pp. 62-73).

The following instructions (Appendix D) and practice passage with judgment of learning scales (Appendix E) were given to participants at the beginning of the second session. All think-aloud sessions were recorded on digital voice recorders. The recorded files were uploaded to a computer and each session was transcribed. The coding scheme for the think-aloud data was based on Pressley and Afflerbach's (1995) detailed lists of monitoring processes (Appendix F). The coded data were analyzed for both frequencies

and patterns of metacognitive monitoring and evaluation processes readers engaged in while encountering unknown and partially known words in context.

Procedures

The current study utilized a quantitative design and data analysis. The study was explained to students during class time, according to IRB procedures, and students who wished to participate signed up for a time and date to complete session one in the laboratory. Reading comprehension and vocabulary subscales of the Nelson-Denny Reading Test and word-knowledge pretest were counterbalanced across participants and took approximately 30 minutes to complete by paper and pencil. At the end of the session participants signed up for a time and date to complete session two.

One week later participants were administered session two measures by computer in the laboratory. By allowing one week between sessions to elapse, participants were likely to have forgotten specific words on the word-knowledge pretest which should have contribute to the validity of word-knowledge posttest data. Second session measures included a demographics form (Appendix G), the four counterbalanced text passages, main idea questions after each passage, judgment of learning scales after each passage, and the word-knowledge posttest with confidence scales. The plain and bolded keywords conditions were counterbalanced to ensure similar sample sizes for analysis. Passages and posttest words were counterbalanced to randomize the potential time elapsed between participants' exposure to a specific word in one of the passages and their generating a response for that word on the posttest. The only addition to procedure for those chosen to be digitally recorded was the think aloud protocol and practice passage.

Data Analyses

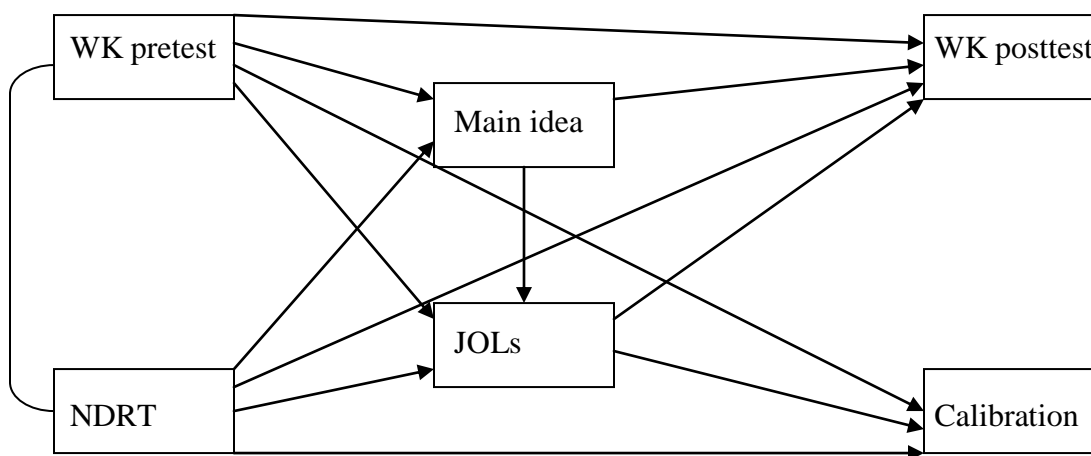
A measured variable path analysis (Figure 2) was utilized to model directional influences of reading subcomponents and metacognitive monitoring on word learning outcomes and calibration of word learning (Research question 1). More specifically, prior word knowledge, general reading skill, understanding of passage main idea, and passage JOLs were hypothesized to influence word learning outcomes and calibration of word learning. Further, the path analysis was used to determine whether the model had appropriate fit for both a plain text (incidental) condition ($n = 154$) and keyword-bolded text (intentional) condition ($n = 155$). Previous literature has indicated that readers have better word learning outcomes and higher calibration of word learning (e.g., Konopak, Sheard, Longman, Lyman, Slaton, & Atkinson, 1987) when readers are presented with an intentional word learning condition (i.e., a manipulation of the text that draws attention to particular words to know or be learned). Examination of path coefficients in the path analysis was used to determine if prior word knowledge is related to JOLs and calibration (Research question 2).

The types of monitoring and strategic processing students employed while reading narrative texts were examined through coded think-aloud data (Research question 3). These data were described in conjunction with main idea, word learning, and calibration outcomes in order to discern which processes were reported by more and less successful readers and word learners. The underlying question to be answered with these data was if indeed successful reading comprehension and successful word learning are overlapping outcomes. It might be the case that paying attention to word level information is too costly for overall comprehension. Think-aloud data were deemed the most appropriate

means to uncover monitoring and strategic processing in an exploratory manner because students could report what they did while reading without being prompted. There is little evidence to suggest whether or not readers pay attention to partially known or unknown words while reading (Shore & Durso, 1990) and therefore a survey or checklist was deemed less appropriate.

Within-subjects repeated measures ANOVAs were run to determine differences in participants' main idea outcomes and JOLs by passage (Research questions 4 and 5). This analysis was chosen because the passages were carefully chosen to vary on several factors of text difficulty.

Figure 2. *Proposed model of word learning and calibration of word learning*



Students' word-knowledge pretest scores were hypothesized to influence their word knowledge posttest scores because the extent of prior word knowledge students demonstrated on the pretest would be somewhat indicative of outcomes on the posttest. Studies have shown that word learning is an incremental process (Durso & Shore, 1991)

and the nature of prior word knowledge affects the amount and quality of new information that can be learned about word meanings. Students' word-knowledge pretest scores were also hypothesized to influence their passage main idea scores because sufficient vocabulary knowledge is a critical component and predictor of reading comprehension (Davis, 1944; Graves, 2000). Since word-knowledge pretest scores were theoretically supposed to have some bearing on passage main idea scores, it was also hypothesized that passage JOLs would be influenced by students' word-knowledge pretest scores. Lastly, students' word-knowledge pretest scores were hypothesized to influence their calibration of word learning indicators because students would use their prior knowledge in assessing their confidence in posttest performance. Additionally word-knowledge pretest scores were hypothesized to correlate with NDRT scores because they each measure vocabulary knowledge.

Students' NDRT scores were hypothesized to influence their passage main idea scores because the NDRT measured general reading comprehension and the main idea passage scores indicated reading comprehension for specific passages. Therefore if students' NDRT scores were low, it was expected that their main idea passage scores might also be low. Students' NDRT scores were hypothesized to influence their passage JOLs because students were thought to have an awareness of their general reading skill and to consider that level of skill when determining their passage JOLs. Previous studies have found that judgments were based on both students' perceived skills and the difficulty of the task (Dinsmore & Parkinson, 2011). Students' word-knowledge posttest scores were hypothesized to be influenced by students' NDRT scores simply because both measures tap into vocabulary knowledge. Students' general vocabulary knowledge,

as measured by the NDRT was thought to have a direct impact on their passage-specific vocabulary knowledge. Lastly, students' calibration of word learning was hypothesized to be influenced by their NDRT scores for the same reason NDRT scores were hypothesized to influence passage JOLs. Students were expected to consider their level of skill as well as the difficulty of the task when making their confidence ratings (Dinsmore & Parkinson, 2011).

Students' passage main idea scores were hypothesized to influence their passage JOLs, as by definition students were judging to what extent they had learned something from the passage. Students' word-knowledge posttest scores were hypothesized to be influenced by their passage main idea scores because students who did not very well understand the passage were supposed to have had difficulty deriving new word meanings from the text. This was determined because students rely on contextual clues from the surrounding passage in order to derive new word meanings (Nagy, Herman, & Anderson, 1987). Therefore, poor comprehension of the passage would limit the amount of information available to derive word meanings. Students' passage main idea scores were not hypothesized to influence their calibration of word learning indicators because passage level performance was deemed to global to influence confidence decisions for individual words that may have been known to some extent before being exposed to the passages.

Students' passage JOLs were hypothesized to influence their word-knowledge posttest scores because if students found the passages to be difficult or confusing they were unlikely to use passage information to help them define words on the posttest, leaving them with only their prior word knowledge on which to rely. Similarly, students'

passage JOLs were hypothesized to influence their calibration of word learning indicators because students who had low JOLs were also expected to have low confidence ratings on word-knowledge posttest items.

Students' word-knowledge posttest scores were not hypothesized to relate to their calibration of word learning indicators because previous studies have found students to be poorly calibrated to cognitive tasks (Glenberg & Epstein, 1985; Lichtenstein & Fischhoff, 1977). The current study sought to determine which factors influence word learning from text and which factors influence calibration of word learning. There was no expectation, based on the existing calibration literature that students' actual outcomes would be related to their calibration indicators. Previous studies of calibration have found that students are not very well-calibrated to reading tasks (e.g., Glenberg & Epstein, 1985). Being poorly calibrated indicates very little relation between actual performance and reported confidence in performance. For this reason it was hypothesized that calibration would have no relation to word learning. Calibration was an outcome in the model because the same person factors hypothesized to relate to word learning were also thought to relate to calibration of word learning.

The paths hypothesized to differ between the two conditions were the path from main idea to word-knowledge posttest, the path from JOLs to word-knowledge posttest, and the path from JOLs to calibration. These paths were hypothesized to differ because participants would focus on different aspects of the passages based on the condition they were assigned (i.e., the instructions they were given). Participants who were instructed to read the passage carefully with no further instruction were expected to focus on the overall meaning of the passage (reading comprehension) without paying much attention

to the target words. Participants who were instructed to read the passage carefully to understand the meaning of the passage and the bolded words were expected to direct their efforts towards deriving word meanings for the bolded words. This difference in focus was hypothesized to have an effect on the main idea statements generated and the JOLs made for the passages. These differences in passage comprehension and judgments of learning from the passages were in turn expected to have somewhat different relations to the word-knowledge posttest and calibration across conditions.

CHAPTER IV

RESULTS AND DISCUSSION

The purpose of this study was to model key factors that influence word learning from text and also to determine the differential effects of text difficulty on reading comprehension and judgments of learning (JOLs) for the specific passages. To this end, I examined how prior word knowledge, reading skill, passage comprehension, and passage JOLs influenced students' word learning outcomes and calibration of word learning. Descriptive self-report was explored in order to make inferences about monitoring and strategic processing during word learning from text. I also examined whether text difficulty had an effect on passage comprehension and passage JOLs. Measured variable path analysis and repeated measures ANOVAs were used in order to analyze data to answer the research questions posited for this study. This chapter consists of three major sections describing the results of analyses for the three corresponding research questions.

Before beginning the path analysis and repeated measures ANOVAs, analyses were run to ensure that data met requisite assumptions. Univariate skewness (pretest = .54, NDRT = -1.34, main idea = .54, JOL = -0.79, posttest = .46) and kurtosis (pretest = .53, NDRT = 3, main idea = -0.1, JOL = 0.73, posttest = .19) of the variables did not exceed an absolute value of 3 (Finney & DiStefano, 2006). As an indicator of multivariate kurtosis, Mardia-based $\kappa = .3$. Calibration was calculated using the formula for Spearman's rho to correlate scored word-knowledge posttest responses with reported confidence ratings, $M_\rho = 0.36$, range: -0.61 to 0.84. This method was chosen because assigning ranks allowed pairing of the variables even though they were on different scales. Selected scatterplots of participants' rho are presented in Appendix J to illustrate

no calibration, very poor calibration, very high calibration, and a mid-level of calibration. Bias (mean confidence score – percent correct on the posttest) was calculated to determine if participants tended to be over- or under-confident in their ratings. In other words, bias provides a sense of directionality for the poor accuracy indicated by the calibration scores. The sample was somewhat overconfident as a whole ($M = 19.78$, $SD = 14.66$) with a wide range of bias from -38.33 to 58.18.

Multigroup Analysis of Incidental and Intentional Word Learning

The first research question addressed whether the proposed model of word learning and calibration of word learning exhibited better fit for the incidental word learning or intentional word learning conditions. Participants were assigned to either a plain text condition with instructions to read the passages carefully; or a bolded target words condition with instructions to read the passages carefully to understand both the overall meaning and the meaning of the bolded words. If participants learned new word meanings from the plain text condition incidental word learning was hypothesized to have occurred. If participants learned new word meanings from the bolded target words condition intentional word learning was hypothesized to occur. Recall that the main theoretical distinction between these two types of word learning was that intentional word learning entails a level of awareness and conscious effort to learn the target words, while intentional word learning happens almost as a byproduct of comprehending what one is reading. Means and standard deviations for each measure included in the model are displayed for both the plain text and bolded target words conditions in Table 5.

Less than 10% of the confidence scale and word-knowledge posttest data were missing at random, generally consisting of a few skipped responses due to technical or

user error. Missing data were estimated using expectation maximization in PASW 18 (Allison, 2002; Enders, 2006).

As shown in Table 5, participants in both plain text and bolded keywords conditions gained some word knowledge and demonstrated poor calibration of their word learning. Using the conditions as a grouping variable, a multigroup measured variable path analysis was conducted in order to determine if the proposed theoretical model exhibited better goodness of fit for the plain text or bolded keywords condition. This analysis was run using maximum-likelihood estimation in EQS (Bentler, 1998) and followed the steps outlined by Kline (2005) for testing multigroup models. First, each group was tested separately for model fit; second, groups were tested simultaneously for model fit; third, a constrained model of the simultaneous groups was tested; and fourth, a constraint was chosen to be released based on the results of the Lagrange Multiplier Test and the model was tested again. Eleven cases contained data missing at random from the remaining variables not estimated using expectation maximization in PASW 18: word-knowledge pretest, Nelson-Denny Reading Test, main idea, or JOL. These missing data were estimated using full information maximum-likelihood in EQS (Bentler, 1998).

Table 5

Means and Standard Deviations for All Variables in the Path Analysis

Variable	Max.	<u>Plain Text</u>		<u>Bolded Target Words</u>	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
WK	100	27.18	12.14	28.57	12.12
pretest					
NDRT	156	129.25	13.87	130.31	12.76
Main idea	4	1.47	1.00	1.40	1.06
JOL	100	71.21	15.83	73.88	14.56
WK	100	31.40	12.58	34.38	12.36
posttest					
Calibration	1	0.37	0.25	0.35	0.23

Note. Plain text condition ($n = 154$); Bolded target words condition ($n = 155$).

Table 6

Covariances for All Variables in the Path Analysis

Variable	1	2	3	4	5	6
1. WK pretest	--	111.36	4.83	88.38	125.84	-0.67
2. NDRT	100.69	--	6.67	92.50	118.83	0.01
3. Main idea	5.97	6.86	--	6.01	5.83	0.00
4. JOL	87.76	84.92	6.95	--	93.09	-0.34
5. WK posttest	121.15	111.38	7.38	100.48	--	-0.78
6. Calibration	-0.86	-0.42	-0.05	-0.95	-0.78	--

Note. Plain text condition is above the diagonal and bolded keywords condition is below the diagonal.

In order to test the model fit of each condition, a multigroup measured variable path analysis using maximum likelihood analysis was run in EQS (Bentler, 1998) in which constraints were imposed to force the paths across groups to be equivalent. Model fit was determined using three indices recommended by Hu and Bentler (1999). The comparative fit index (CFI) compares the hypothesized model to the null model and has been recommended to be greater than or equal to .95 to retain the model. The standardized root-mean-square residual (SRMR) is an absolute fit index that averages values in the residual covariance matrix and has been recommended to be less than or equal to .08 to retain the model. Finally the root-mean-square error approximation (RMSEA) indicates the parsimony of the hypothesized model and has been recommended to be less than or equal to .06 to retain the model.

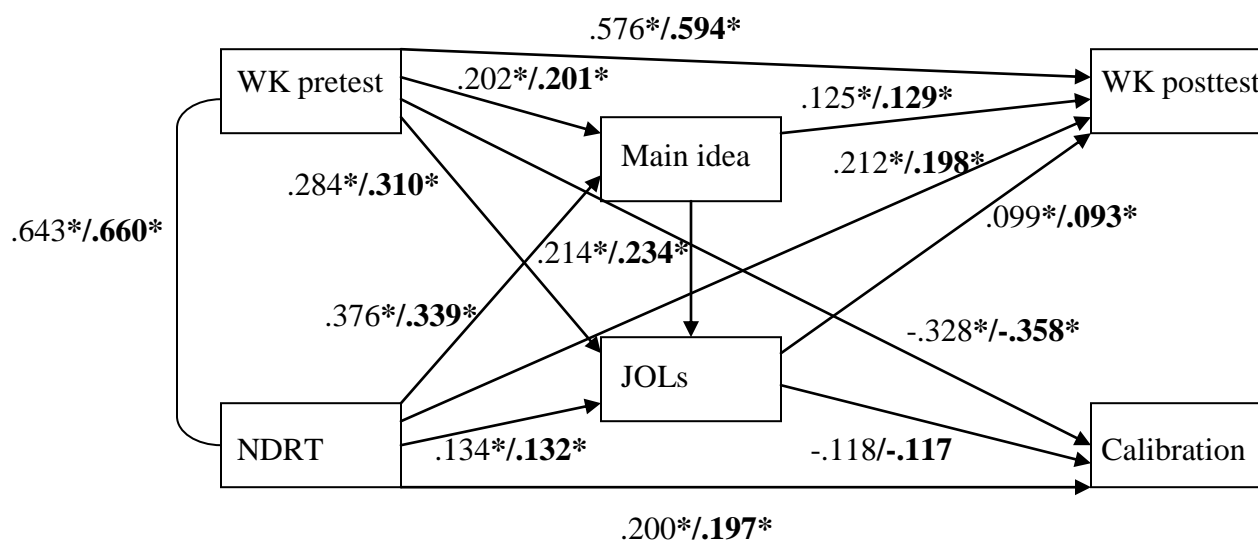
The simultaneous multigroup model with constraints had excellent fit, CFI = 1, SRMR = 0.05, and RMSEA = 0. As shown in Table 7, the overall model fit improved with the constraints imposed, but did not differ significantly when releasing the constraint suggested by the Lagrange Multiplier Test, $\chi^2_{diff}(1) = 2.56$, $p = .15$. This suggests that there are no significant differences in the paths across groups. In other words, the path model is equivalent for the plain text condition and the bolded keywords condition. The model with standardized path coefficients is displayed in Figure 3. Note that although the unstandardized paths were constrained across groups the standardized paths may still vary. Thus, standardized paths were reported for both groups.

Table 7

Summary of Data Model Fit Indices

Model	X^2	<i>df</i>	CFI	SRMR	RMSEA	90% CI (RMSEA)
Simultaneous groups	1.28	4	1	0.01	0	(0, 0.15)
Constraints across groups	8.29	14	1	0.05	0	(0, 0.04)
Released constraint	5.73	13	1	0.04	0	(0, 0.02)

Figure 3. Model with Standardized Path Coefficients



Note. Standardized path coefficients from the plain text condition appear in plain text and standardized paths from the bolded keywords condition appear in bold. Path coefficients designated with an * were significant at $p < .05$.

The largest effect sizes were seen in the influence of word-knowledge pretest score on word-knowledge posttest score (.576 and .594 for the plain text and bolded keywords conditions respectively). This path is theoretically important because it suggests that students' prior vocabulary knowledge is a key predictor in their resulting word-learning outcome after reading a text. Previous research has suggested that students must be taught key word meanings before reading assigned texts or stories in order to effectively comprehend the text and learn necessary meanings or concepts (Graves, 2000).

The influence of Nelson-Denny Reading Test scores on passage main idea scores also had a large effect size (.376 and .339), suggesting that general reading skill accounted for a portion of students' ability to comprehend the main ideas of the passages.

Lastly, there was a relatively large negative effect of word-knowledge pretest score on calibration of word learning (-.328 and -.358). This path will be further discussed in the following section.

The Influence of Word-Knowledge Pretest Scores on JOLs and Calibration

In order to more fully explicate the influence of word-knowledge pretest scores on JOLs, the direct and indirect effects were summed for a total effect of .413 and .444 for the plain text and bolded keywords conditions, respectively. Total effects and variances were reported separately for each condition because they were calculated from the standardized path coefficients. Word-knowledge pretest scores accounted for 17.06% and 19.71% of the total variance in participants' JOLs, perhaps suggesting prior word knowledge was one factor participants considered when making their judgments about how well they learned from the passages. The more word knowledge students demonstrated on the pretest, the higher their JOLs tended to be.

This positive relation was not echoed in the relation between word-knowledge pretest scores and calibration of word learning. The total effect was .496 and .530 for each condition. This effect was negative in direction, meaning the higher participants scored on the word-knowledge pretest the lower their calibration indicator (i.e., rho). This suggests that those participants who had low scores on the word-knowledge pretest were aware of their lack of word knowledge and were more accurate in their assessment of their word learning performance on the word-knowledge posttest. Word-knowledge pretest scores accounted for 24.60% and 28.09% of the total variance in calibration of word learning indicators.

Processes and Strategies Related to Word Learning and Calibration of Word Learning

Think aloud data were collected from a subset of participants in order to investigate the third research question. Those who participated in the think aloud indicated that they were willing to be audio recorded and were on the schedule at a time when they could have a whole room to themselves. Twelve participants from the study met these criteria (with the private room being the main limiting factor as several students were generally participating in the study simultaneously). Eleven of those 12 participants produced audio recordings that could be coded for processes and strategies. The participant that was excluded from coding did not make any utterances while reading.

Overall, there were no discernible patterns in measure outcomes between the five participants who received the plain text condition and the six participants who received the bolded keywords condition, similar to the larger sample. Discussion of the cases will begin with those participants who showed the largest net gains in word knowledge from pretest to posttest and then those who showed the largest net losses. Then those who were most highly calibrated will be described, followed by those who were most poorly calibrated. Finally, a participant who was about average for the subsample on both change in word knowledge and calibration will be examined. When participants are characterized as high or low in a particular category it refers to their being at least one standard deviation above or below the subsample mean for that category.

Strategic Processes Reported and Highest Word Knowledge Gains

Jean had an overall gain of 22.5 points from her word-knowledge pretest to her word-knowledge posttest. Her scores on the NDRT (131) and word-knowledge pretest

(51) were average for the think-aloud subsample. She also reported using an average number (27) and variety (6) of strategies while reading the passages. Jean used strategies such as local restatement and interpretation while reading. She mentioned feeling challenged three times, monitoring once, and referenced her knowledge four times (all average frequencies). Her main idea score (3) was high for the subsample and her main idea statements were both accurate and detailed. Jean was unique in gaining knowledge for 10 of the target words and not losing knowledge of any of the other words. She correctly used *banter* in one of her main idea statements and also demonstrated increased knowledge of the meaning from pretest to posttest. It seems that comprehending the passages may have contributed to Jean's word learning. She demonstrated little to no calibration of her word learning ($\rho = .16$) and completed the bolded keywords condition.

Molly also showed a high overall gain in word knowledge from pretest to posttest (13.5 points). As with Jean, she had an average NDRT score (135) and average word-knowledge pretest score (36). Unlike Jean, however, she reported using a low number (6) and variety (3) of strategies while reading and made no mention of challenge or monitoring (also low). She did mention situational interest once. Also unlike Jean, Molly's main idea score was low (0 points) and her statements were rambling and did not make much sense. She gained knowledge for seven words from pretest to posttest and lost knowledge for four words. Based on Molly's outcomes reading comprehension seemed to have much less relation to her word learning than it did for Jean. Molly's level of calibration was average for the subsample ($\rho = .28$) and she completed the plain text condition.

Strategic Processes and Largest Decline in Word Knowledge

Lucy lost the most overall points from word-knowledge pretest to posttest (-7). Like Jean and Molly she had an average NDRT score (122) and an average word-knowledge pretest score (21). She reported using an average number (23) and variety (6) of strategies while reading. Lucy mentioned a high number of feelings of challenge (13) and a high frequency of monitoring (6). When Lucy expressed challenge she made statements such as, "I don't know because it's so confusing," or "I understood everything up until then." She also mentioned situational disinterest with high frequency (7), with statements such as "These are really weird paragraphs." Her main idea score was average (1.5) with partially accurate statements containing no miscues. Despite Lucy's overall decline in word-knowledge performance from pretest to posttest, she did gain knowledge for two words. Lucy differs most from Jean and Molly in her statements of challenge and disinterest while reading. This frustration may have negatively impacted her word learning since she seemed to confuse the meanings of some words at the posttest that she knew something about at the pretest. Lucy did not demonstrate much calibration ($\rho = .11$) and she completed the bolded keywords condition.

Linus had the second largest decline in performance pretest to posttest (-3). Like the others, his NDRT score (133) and word-knowledge pretest score (34.5) was average. He reported using a high number (49) and variety (9) of strategies while reading and, like Lucy, reported a high frequency of challenge (17) and monitoring (8). Linus used strategies such as rereading and elaboration. Unlike Lucy, Linus also frequently referenced his knowledge (12) and expressed situational interest (9). His strategic efforts were reflected in a high main idea score (3) and his statements were extremely accurate

with the exception of the most difficult text. Despite his success comprehending the passages, Linus had somewhat mixed results in his word learning. He gained new knowledge for four of the words, but decreased in performance for five of the words from pretest to posttest. Linus was also the most highly calibrated word learner in the subsample ($\rho = .51$). He completed the bolded keywords condition.

Strategic Processes and the Highest Levels of Calibration

Although Linus had one of the largest decreases in overall word-knowledge performance of the subsample, he was the most highly calibrated individual in the group. He also reported the most extensive strategy use and the most frequent monitoring and situational interest. All of these processes suggested a student who persistently self-regulated his learning from the text and was well aware of his level of performance on the word-knowledge posttest.

Belle was nearly as well calibrated as Linus (.49). While she had an average NDRT score (140), her word-knowledge pretest score was high (60). She used an average number (29) and variety (7) of strategies and mentioned challenge (3), references to her knowledge (2), and situational interest (2). Unlike Linus, Belle did not report monitoring at all while she was reading. Her main ideas were of variable accuracy (2.5), but she did use target words in her main idea statements. For example, according to Belle's pretest she knew something of the meaning of *dilapidation* and used it correctly in her main idea statement for the passage in which it appeared. However, Belle's definition of *dilapidation* was not as accurate on the posttest as it was on the pretest. Overall, she gained 4.5 points from pretest to posttest. She was fairly well calibrated to

her word learning ($\rho = .49$) despite this shifting so she seemed to be somewhat aware of her changing understandings. Belle completed the bolded keywords condition.

Strategic Processes and the Lowest Levels of Calibration

Marcie was the most poorly calibrated of the participants in the subsample ($\rho = -.16$). She had a high NDRT score (151) and a high word-knowledge pretest score (63). She reported an average number of strategies (9), but each strategy was different so her variety of strategies was high for the subsample. She made average mention of challenge (1), monitoring (2), reference to prior knowledge (1), and situational interest (1). Her main idea score was also average (2.5) and her statements were long, but not always accurate. Although Marcie showed no overall change in word knowledge from pretest to posttest she improved her performance for four individual words and decreased in her performance for four individual words. Marcie had high general reading skill and high prior knowledge for the target words, yet she was poorly calibrated to her performance on the word-knowledge posttest.

Lucy was the second most poorly calibrated participant in the subsample ($\rho = .11$). As described above, Lucy did not have as high a skill level or prior word knowledge as Marcie did and she also indicated a high level of challenge and disinterest. Lucy's performance declined overall from pretest to posttest (-7). While Marcie was poorly calibrated, as indicated by the negative rho value, Lucy demonstrated very little calibration, as indicated by a rho close to zero. Although they were the worst calibrated individuals in the subsample, Marcie was clearly more poorly calibrated than Lucy.

Table 8
Think aloud data

Name	NDRT	WKpre	# of strategies	Variety	Challenge	Monitor	Ref know	Interest	Main Idea	WK post	Calib
Group	130(16.35)	28.6(13)	20.1(12.58)	5.91(2.39)	5(5.66)	2.64(2.77)	2.55(3.5)	.27(3.13)	1.73(.93)	33.73(13.3)	.28(.20)
Sally	83	6.5	21	6	7	6	0	-1	2	19	.35
Linus	133	34.5	49	9	17	8	12	5	3	31.5	.51
Marcie	151	54	9	9	1	2	1	1	2.5	54	-.16
Frieda	136	27.5	4	2	2	2	0	-1	2.5	32.5	.18
Violet	134	28	18	6	8	2	3	-1	.5	27	.42
Peggy	134	27	20	8	1	2	2	4	1.5	33	.35
Patty	127	20	15	3	0	0	0	0	1.5	24	.38
Lucy	122	21	23	6	13	6	2	-7	1	14	.11
Jean	131	29	23	6	13	6	2	-7	1	51.5	.16
Belle	140	46.5	29	7	3	0	2	2	2.5	51	.49
Molly	135	20	6	3	0	0	0	1	0	33.5	.28

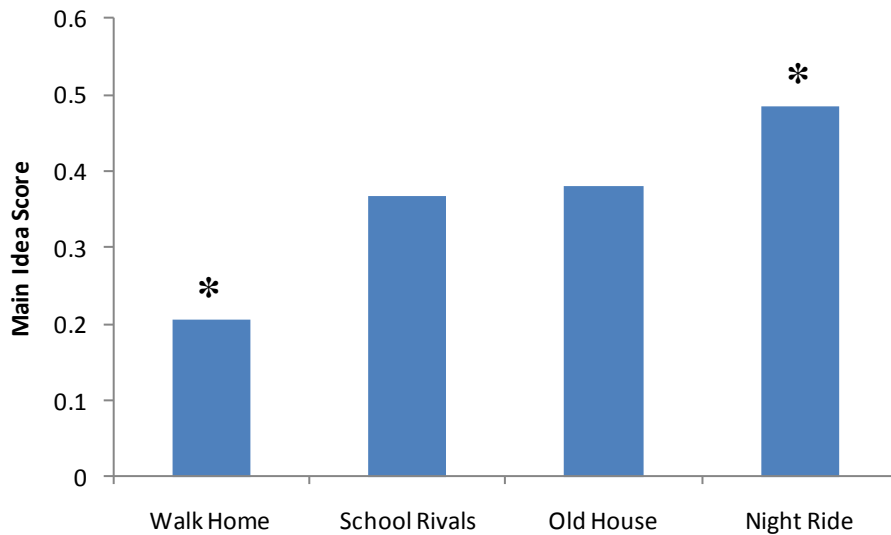
Effect of Passage Difficulty on Main Idea and JOL Scores

The fourth and fifth research questions examined text features rather than person factors in interpreting main idea scores and JOL ratings. The passages were chosen to be globally more or less difficult based on the Lexile Framework (Lennon & Burdick, 2004). The Lexile rating also gave support to the claim that these texts are appropriate for college undergraduates as the Lexiles are designed to indicate grade level difficulty. Repeated measures ANOVAs were run in order to determine if differences in main idea score and JOLs could be detected across passages.

Differences in Readers' Comprehension

A within-subjects repeated measures ANOVA was run to determine differences in participants' main idea outcomes by passage. Mauchly's Test of Sphericity for this analysis was significant, Mauchly's $W = .95, p < .01$. Since sphericity could not be assumed the Huynh-Feldt adjustment was used, $F(2.93,305) = 28.97, p < .01, \eta^2 = .086$. Tukey HSD pairwise comparisons were examined to determine where differences existed between passages (Figure 4). Participants scored significantly lower for the Irving passage *Walking Home* ($M = .21, SD = .36, p < .01$) than the other three passages. Conversely participants scored significantly higher for the Irving passage *Night Ride* ($M = .49, SD = .45, p < .01$) than the three other passages. The two Poe passages, *School Rivals* ($M = .37, SD = .43$) and *Old House* ($M = .38, SD = .41$) were not significantly different from each other ($p > .10$), however they were each different from the two Irving passages ($p < .01$ in both cases).

Figure 4. Mean differences in main idea scores by passage

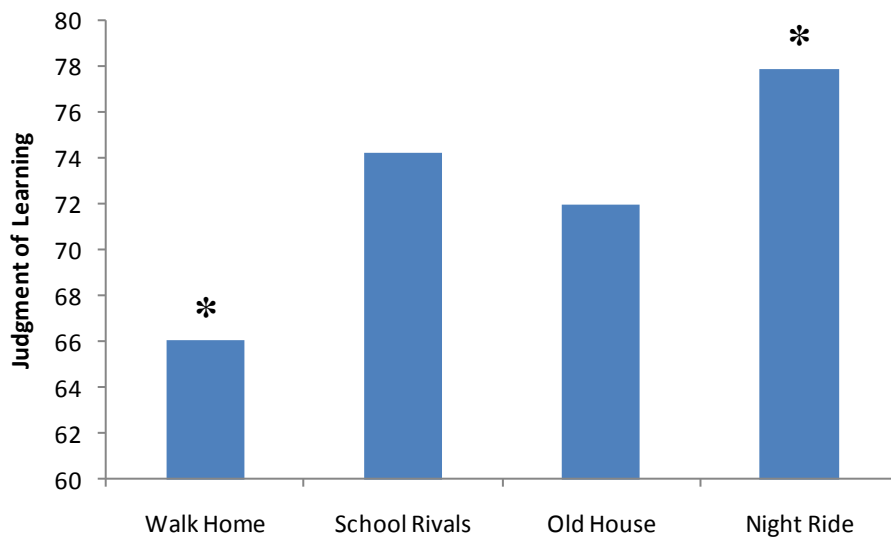


Differences in Readers' Judgments of Learning

Another within-subjects repeated measures ANOVA was run to determine differences in participants' judgments of learning by passage, $F(3,305) = 31.97, p < .01, \eta^2 = .10$. Mauchley's Test of Sphericity was not significant for this analysis and therefore sphericity was assumed. Tukey HSD pairwise comparisons (Figure 5) revealed that participants had significantly lower JOLs for the Irving passage *Walking Home* ($M = 66.06, SD = 22.25, p < .01$) than the other three passages. Participants had significantly higher JOLs for the Irving passage *Night Ride* ($M = 77.86, SD = 19.93, p < .01$) than the other three passages. The two Poe passages, *School Rivals* ($M = 74.18, SD = 19.32$) and *Old House* ($M = 71.95, SD = 21.83$) were not

significantly different from each other ($p > .05$), however they were each different from the two Irving passages ($p < .01$ in both cases).

Figure 5. Mean differences in JOLs by passage



Text-level and word-level factors affecting these observed differences

A traditional text readability index that takes into account sentence length and word frequency (The Lexile Framework) rated the difficulty of both Irving passages (*Walk Home* and *Night Ride*) as the same. However, this cannot account for observed differences in main idea and JOL outcomes. A more thorough investigation into text factors affecting these outcomes was deemed necessary. These two passages were analyzed using Coh-metrix 2.1 (McNamara, Louwerse, Cai, & Graesser, 2005). *Walk Home* was found to have higher causal cohesion (8) than *Night Ride* (.5). Causal cohesion was lower in *Night Ride* because there were fewer causal particles to connect subjects to verbs than there were causal verbs. The results for causal

cohesion were unexpected as participants were better able to state the main idea for *Night Ride* than they were for *Walk Home*. Since causal cohesion did not account for these differences two other indices of text complexity were examined.

Walk Home had a higher frequency of adjacent anaphoric reference (.75) than did *Night Ride* (.625). An anaphoric reference is when a pronoun refers to the subject(s) of previous sentences. The greater frequency of anaphoric references in *Walk Home* indicates that this passage may have been more difficult to understand compared to *Night Ride*. Lastly, *Walk Home* contained a higher frequency of modifiers per noun phrase (1.066) than *Night Ride* (.813). Increasing the number of adjectives before a noun adds complexity to the syntactic complexity of sentences, thereby making them more difficult to comprehend. These indicators suggest that the Lexile framework alone could not account for differences in text difficulty.

CHAPTER V

SUMMARY, IMPLICATIONS, AND FUTURE RESEARCH

Summary and Conclusions

This study was designed to address several gaps in the existing literature on word learning. Two of the most important issues addressed were the lack of study on word learning during reading (as opposed to during artificially designed tasks) and the lack of information about how prior word knowledge, reading skill, metacognitive monitoring, and passage comprehension influence word learning. This summary will draw conclusions about the way in which the results reported in Chapter IV can contribute to our understanding of these issues and also knowledge about word learning during reading. Limitations to the study and future research will also be discussed.

Incidental versus Intentional Word Learning

Literature on word learning has made a distinction between incidental word learning, where the purpose of word learning during reading was not purposefully evoked, and intentional word learning, where students were directed to derive word meaning from text (Swanborn & de Glopper, 1999). A recent meta-analysis on incidental word learning only analyzed 15 studies of monolingual word learners (Swanborn & de Glopper, 1999), drawing attention to the fact that intentional word learning studies have been overwhelmingly more extensively studied. The purpose of the meta-analysis was to draw attention to the distinction between incidental and intentional word learning and urge researchers to consider the importance of investigating incidental word learning.

An empirical question in the current study was whether or not incidental word learning conditions could be considered to be fundamentally different from intentional word learning conditions. Fit indices from the current study indicated that there was no difference in model fit across the incidental and intentional word learning conditions. This finding provides a piece of evidence indicating that incidental and intentional word learning conditions might not have differential effects on word learning for undergraduates reading narrative texts. Perhaps students were primarily concerned with comprehending the stories contained in the passages regardless of the different directions given for each condition. This would support Daalen-Kapteijns and Elshout-Mohr's (1981) notion that typically readers only stop to take notice of unfamiliar words when the words impede comprehending the text.

The Influence of Prior Word Knowledge on Metacognitive Monitoring and Calibration

Previous research on expertise and metacognitive judgments, such as judgments of learning and confidence ratings, has indicated that metacognitive judgments appear unaffected by expertise (Lichtenstein & Fischhoff, 1977). In other words, having more knowledge did not affect the accuracy of confidence judgments. However, there have not been any studies on the influence of metacognitive judgments on word learning, so this became the second empirical question. Prior word knowledge, as assessed by the word-knowledge pretest, was found to positively influence metacognitive monitoring (i.e., JOLs). More specifically the path model indicated that prior word knowledge had a direct effect on JOLs, suggesting that students may have recognized the necessity of understanding particular words in

order to assess their learning or comprehension from the text. The finding in the current study that increased prior knowledge did contribute to increased JOLs may have differed from Lichtenstein and Fischhoff's (1977) findings simply based on their operationalization of expertise. Students in the current study were not assumed to be expert readers, however, they were assumed to have extensive experience in reading texts and were supposed to be capable of being reflective about their understanding of a particular text. Perhaps Lichtenstein and Fischhoff failed to find a relation between expertise and monitoring accuracy because they did not provide tasks with which participants were actually all that familiar.

The path model also indicated that calibration of word learning was negatively influenced by prior word knowledge. In other words, the more prior word knowledge participants demonstrated, the more poorly calibrated they were to their performance on the word-knowledge posttest. This was surprising because prior word knowledge was hypothesized to positively influence calibration of word learning since students who knew more meanings before seeing the words in context were expected to have a better basis on which to judge their understanding of word meanings after seeing the words in context. As this was not the case, prior word knowledge cannot be considered a source of information students appropriately use when judging their performance confidence. Perhaps the more students knew on the word-knowledge pretest the better they expected to perform on the word-knowledge posttest, thus inflating their confidence ratings compared to actual performance. Overall, students were overconfident relative to their actual demonstration of word knowledge,

however there was wide variation in the degree of overconfidence or underconfidence between participants.

Processes Reported During Reading

Think-aloud protocol has been a popular methodology to study reading comprehension (Pressley & Afflerbach, 1995) and word learning during reading (Fukkink, 2005) since it serves to uncover readers' strategies and processes. The purpose of think-aloud data in the current study was to illuminate some of the underlying processes that support word learning during reading and calibration of word learning. In other words, if students struggled to comprehend the main idea of a passage would they be able to derive word meaning from that confusing passage?

Results from the coded think-aloud transcripts suggested that there was not an ideal pattern of strategic or skillful processing that led to good reading comprehension performance, large amounts of word learning, or high levels of calibration of word learning. Rather the role of strategic and skillful processing in supporting word learning during reading and its subsequent relation to calibration of word learning varied across individuals. Those students who seemed to gain the most word knowledge from pretest to posttest demonstrated a wide variability in their passage comprehension performance, and reported using varying amounts and variety of strategies while reading.

Those students who decreased the most in their word-knowledge performance from pretest to posttest were those who most frequently reported feeling challenged while reading the passages. Perhaps this created a sense of frustration and confusion that interfered with their ability to provide coherent or accurate definition on the

word-knowledge posttest. Students who demonstrated the highest levels of calibration of word learning tended to do well on passage comprehension measures, but they varied widely in their self-reporting of monitoring while reading. This supports the finding in the larger sample that metacognitive monitoring (i.e., JOLs) did not influence calibration of word learning. One reason for this may have been that because JOLs were passage level judgments they were not specific enough to relate to the finer-grained confidence questions about individual words. Dunlosky, Rawson, and Middleton (2005) found that judgment accuracy was better for term-specific judgments than for passage-level judgments because the passage level judgments were too global in nature.

Students who were the most poorly calibrated seemed to use an average number of strategies and had average passage comprehension performance. There was nothing to distinguish their strategic processing from those students who demonstrated better levels of calibration. This suggests that poor calibration of word learning cannot be blamed on lack of strategic processing during reading nor lack of passage comprehension. For these students poor calibration was unrelated to their general reading skill or prior knowledge as well. It is plausible that these students were overconfident in their word-knowledge posttest performance because they gave no indication through their think-aloud utterances that they were experiencing difficulties in comprehending or understanding parts of the text containing the target words. Due to these students' overconfidence in their word knowledge they would be less likely to seek help (in this case by consulting a dictionary) in order to correct misunderstandings or false attributes of word meaning. In the college classroom this

would translate into students not seeking further clarification when they failed to recognize that they did not fully understand particular concepts.

The main conclusion to be drawn from the think-aloud data was that students whose word-knowledge performance decreased from pretest to posttest were aware of the challenges they were facing comprehending the passages and understanding the difficult words. Students whose word-knowledge performance increased from pretest to posttest did not have discernible similarities in their strategic processing or passage comprehension scores. To this end the think-aloud data captured did not support a clear connection between either strategic processing and reading comprehension or strategic processing and word learning. Seeing the target words in just one context may explain the inability of think-aloud protocol to capture any connection between word learning and strategic processing during reading. The instance-based approach to word learning suggested that multiple exposures were necessary before students were able to integrate multiple meanings of a single word and use those meanings flexibly (Bolger, Balass, Landen, & Perfetti, 2008). If multiple exposures were necessary for students to demonstrate their understanding of word meanings then those students would not be expected to be able to verbalize their efforts to reconcile their prior knowledge with new knowledge provided by context after just one exposure. Awareness of particular words would first have to be raised by their appearance in multiple contexts and then participants would be more likely to self-report their strategic processing in coming to know particular words. Additionally the difficulty of some of the texts in the current study may have drawn all metacognitive

efforts to comprehending the text and left no resources for attending to individual word meanings.

Text Factors Impacting Reading Comprehension and Metacognitive Monitoring

Readability formulae have been used for some time in assigning value judgments of text difficulty and even grade level appropriateness. However, decades of criticism against the over-simplicity of reading formulae have prompted alternative means to analyze complexity of texts (McNamara, Louwerse, Cai, & Graesser, 2005). The importance of these alternative measures of text difficulty was highlighted in the results of the current study. Differences were found in students' passage comprehension and JOLs across passages. These differences were surprising because one of the Washington Irving passages (i.e., *Walk Home*), was related to the lowest passage comprehension and JOLs, while the other Washington Irving passage (i.e., *Night Ride*), was related to the highest passage comprehension and JOLs. According to the Lexile Framework (Lennon & Burdick, 2004), these passages were at the same level of linguistic difficulty.

A closer inspection using Coh-metrix 2.1 (McNamara, Louwerse, Cai, & Graesser, 2005) revealed that *Walk Home* was more difficult than *Night Ride* because the former passage contained more anaphoric references than the latter, requiring readers to make referential connections across sentences, and more modifiers for each noun phrase. Some participants did in fact complain about “the weird writing style” or the “flowery, descriptive language” in their main idea statements. Although the conclusion from these findings is not novel it is critical to reiterate that text factors beyond word frequency and sentence length must be considered when choosing texts

for reading comprehension and word learning studies. Since reading comprehension was defined as an interaction of person and text factors (RAND Reading Study Group, 2002) features of the text should certainly affect reading outcomes. The unique contribution of the current study was to consider factors contributing to text difficulty in naturally occurring texts rather than manipulating texts to remove commonly encountered complexities such as anaphoric reference and noun phrase modifiers.

In summary, the current study provided evidence to address major gaps in the literature on word learning during reading. The nature of incidental and intentional word learning was addressed and found to be non-distinct. Recent examination of the importance of metacognitive monitoring and calibration to performance was extended to the study of word learning. In addition to these person factors, text factors were examined in naturally occurring texts rather than manipulated sentences presented in isolation in order to better understand word learning during typical reading. Lastly, these issues were all addressed with a sample of undergraduates who were purposefully chosen due to the necessity of word learning to their academic success. Undergraduates must rely on reading at least somewhat technical texts in their fields of specialization and also in fields about which they might know very little. The implications of conclusions drawn from the current study have ramifications for research and educational practice.

Limitations

Although the current study made significant contributions to the study of word learning during reading there were a number of limitations to consider in

interpretation of the results. First, the evidence for equivalence between incidental and intentional word learning can only be assumed for undergraduates reading narrative texts. Additionally other factors not included in the model might create differences in incidental and intentional word learning conditions not accounted for in the current study. These limitations could be easily addressed in follow-up studies of different age groups, different genres of text, and with measures tapping different aspects of reading, metacognition, and word learning.

Most importantly, considering motivational or affective factors in future studies would provide a great deal of information about how students approach reading tasks in general. Knowing more about willingness or lack thereof to spend time and effort reading challenging texts would hypothetically be related to variables under investigation in the current study such as judgments of learning, passage comprehension, word learning outcomes, and calibration of word learning. Specifically it would be helpful to know students' goals for reading assigned texts (especially if word learning is among those goals), their interest in the topic and reading in general, and their perceived value for the task.

Another limitation to the current study was that students were not asked what information they used to make their judgments of learning and confidence ratings. A recent study found that students report a number of different factors considered when making confidence ratings (Dinsmore & Parkinson, 2011). Having information about how students came to make their judgments would allow for modification to the current model related to calibration of word learning to more accurately reflect what students claim to be considering.

The current study was not well-designed to collect think-aloud data in order to uncover processes related to word learning during reading. The quality of think-aloud data related to word learning might have been improved by choosing texts that were easier for students to comprehend. Easier texts would have allowed more metacognitive resources to be devoted to challenging or unknown words while reading. Another way to boost utterance related specifically to word learning would have been to present the target words in multiple contexts. Perhaps this would have prompted students to notice particular words and the fact that they were not very familiar with the meanings for those words. If this awareness was made more salient by repeated encounters with words students may have been more likely to self-report their noticing and their efforts to reconcile their varying exposures to the same words.

Implications and Significance

Research

Results of the current study indicated that the theoretical distinction made between incidental and intentional word learning may be somewhat misguided in the case of undergraduates reading narrative texts. Previous research on reading comprehension has found that the nature of directions influences students' purpose for reading (McCrudden, Magliano, & Schraw, 2010). However directions did not seem to affect word learning outcomes in the same manner in the current investigation. A limitation to this comparison between direction for reading comprehension and directions for paying attention to words within text is that it may be easier to change the more global purpose of reading comprehension than to change students' intentions towards word learning while reading. For example, Daalen-

Kapteijs and Elshout-Mohr (1981) theorized that few readers engage with a text specifically to gain new word knowledge. These researchers thought it was somewhat more likely that readers might be aware of words that they were unfamiliar with while reading. But, the most likely case they proposed was that students read solely to understand the overall main idea and supporting details of the text and do not spend effort learning words unless it is necessary for reading comprehension. If indeed it is more difficult to change readers' purpose for word learning from incidental (learning almost by accident) to more intentional this would explain the lack of difference found in model fit for the two conditions.

A second key finding from the current study was that prior word knowledge does relate to metacognitive monitoring and calibration of word learning. Therefore studies of word learning should consider that what students already know affects their metacognitive monitoring during reading and their calibration of word learning after reading. This is especially important because students' partial understandings of word meanings are a kind of fragile knowledge (Alexander, 2004) that requires accurate awareness in order to add knowledge from multiple instances and change to a fully known word meaning (Bolger, Balass, Landen, & Perfetti, 2008).

In order to further investigate this necessary awareness studies may implement think-aloud protocol to capture monitoring and awareness during reading. Since the current study did not find enough information in think-aloud data to support a clear connection between strategic or skillful processing, reading comprehension, and word learning the nature of this connection (or lack thereof) remains an empirical question to be studied. Again, there might be differences in undergraduates' comfort and

willingness to fully engage with expository texts rather than narratives that suppressed the expression of certain strategies. There were many more mentions of situational disinterest while reading than there were of situational interest (Table 7).

Finally, person factors cannot be considered in isolation when approaching word learning during reading because reading comprehension entails an interaction between reader and text (RAND Reading Study Group, 2002). Varying complex factors within texts might account for text difficulty and cannot be captured by a readability formula. While naturally-occurring texts should be used for generalizability to typical reading, a careful analysis of text characteristics is necessary in order to fully understand reading comprehension or word learning outcomes. In the current study, it is possible that word learning was extremely challenging due to the difficulty of at least one of the passages.

Educational Practice

College students frequently encounter unknown vocabulary within texts that are critical for academic success. It is assumed that undergraduates are practiced in the requisite metacognitive processes to learn new words by the time they reach college. Unfortunately, students often use inefficient or ineffective processes to acquire and evaluate knowledge gained from text (Glenberg & Epstein, 1985; Lin & Zabucky, 1998). This leads them to forget or confuse terms related to specific concepts, processes, and main ideas. Increased understanding of how to increase undergraduates' awareness of metacognitive processes for word learning would improve general reading comprehension and classroom learning. Before any such instructional suggestions can be made more must be known about what influences

calibration of word learning. The current study offered evidence that prior word knowledge and general reading skill influence calibration of word learning, but metacognitive monitoring does not.

This study offered several unique contributions to the study of word learning in context. Intentional word learning was directly compared to incidental word learning. Several reviews have addressed these topics individually (see Fukkink & de Glopper, 2001 and Swanborn & de Glopper, 1999), but few studies have directly compared these two conditions.

The significance of better understanding how adult readers calibrate their word learning has several implications. First, understanding the accuracy with which students calibrate their word learning provides insight into potential road blocks to the learning process. Poorly calibrated word learners do not necessarily recognize their level of word knowledge, and therefore do not adjust their reading and learning strategies accordingly. Understanding the relation between calibration and word learning allows educators to improve vocabulary instruction. Previous studies (Fischhoff & Lichtenstein, 1980) have demonstrated that training improves individual's calibration for a particular task. There is potential for systematic improvement in calibration of word learning through instruction. This kind of improvement would allow students to better gauge their learning from assigned texts.

Future Research

Given the findings from the current study there are a number of avenues for future research on word learning during reading to be explored. First, the findings from the current study on incidental and intentional word learning conditions should

be investigated using expository texts. It is important to know whether undergraduates have the same approach to the genre of their textbooks as they do to narratives. Several students in the current study expressed the oddity of reading narratives since they were so accustomed to reading expository text for their classes. This suggests that genre might play a role in how students interpret directions for incidental versus intentional word learning conditions.

Another avenue for future research would be to present multiple naturally-occurring texts containing a set of target words to be learned. Word learning has been studied in multiple contexts in previous research (e.g., Bolger, Balass, Landen, & Perfetti, 2008; McKeown, 1985), however the multiple contexts were constructed by researchers in order to manipulate specific text features such as context clues. Although the task of finding texts that would contain similar target words might be somewhat challenging, it would more closely simulate the typical word learning contexts encountered by students in their daily pursuits. It is important to note that in the current there was evidence of overall word learning from pretest to posttest after a single exposure to the target words in context. While word learning can be detected after just one contextual exposure, it is difficult to understand the relation of reading comprehension and word learning from text after a single encounter with target words in context.

Yet another critically important empirical question to be investigated is reason for students' inaccuracy when calibrating their word learning. The sample of undergraduates reading short narrative texts was found to be generally overconfident in their judgments relative to their actual performance demonstrating word

knowledge. The direction of inaccuracy matters because it has implications for the way in which students perceive the need for help seeking behaviors such as looking up definitions for unknown words, or asking another person for clarification and discussion. Students who are overconfident in their word knowledge would see less of a need to seek help understanding word meanings because they would be less aware of what they did not know. For narrative texts this might not seem like a major issue, but in expository texts the difficult words tend to represent key concepts to be learned.

Calibration can be improved with feedback or training when the accuracy of word learning is critical to success in educational settings. There has been some evidence that calibration can be improved, at least in the short term, for specific tasks (Huff & Nietfeld, 2009). Future studies should focus on the effectiveness of implementing an explicit program of feedback to notify learners when they are inaccurate and suggest what those learners should specifically pay attention to in order to make more accurate judgments. In the case of word learning, students could be notified when they are overconfident in their word knowledge performance and then receive suggestions for what they might consider when making their judgment. Perhaps prompting students to explain what they think constitutes a good definition and asking if their definition meets those criteria would be one suggestion. Providing feedback to train students how to evaluate their word learning from independent reading has the potential to increase their success as learners across domains.

APPENDICES

Appendix A: Word Knowledge Pretest

Pilot Study

Directions: Write a definition or short description for every word that you can on the list. Please make your definitions as clear as possible so that I know that you understand the meaning of the word. I am not interested in the number of words that you know, so just do your best.

Admonish**Arbitrary****Benefactor****Boding****Boorish****Calsar****Capacious****Clove****Congeniality**

Contrived

Countenance

Derision

Devernal

Dilapidation

Dispelled

Docile

Drallen

Edarthic

Expound

Filigreed

Fissure

Forlorn

Fossern

Gilded

Harbinger

Ineffable

Incipient

Jandelar

Kindle

Lucid

Melancholy

Merriton

Motley

Pacific

Prodigious

Pertinacious

Petulant

Phisteron

Pommel

Psalmody

Sagacious

Sentiment

Specious

Specter

Stave

Sullen

Redistac

Tarn

Thonstan

Thwart

Tumultuous

Varlet

Vehemently

Veritable

Vignette

Waggery

Wane

Wended

Withe

Whim

Directions: Next, go through the list again and place a check mark beside any word that you left blank if you have seen it before or if it is familiar to you, even if you are not quite sure what it means.

Dissertation Study

Directions: Write a definition or short description for every word that you can on the list. Please make your definitions as clear as possible so that I know that you understand the meaning of the word. I am not interested in the number of words that you know, so just do your best.

Admonish**Animosity****Anomalous****Banter****Benefactor****Blundering****Capacious****Clove****Con****Congeniality**

Devernal

Dilapidation

Dispelled

Edarthic

Endeavored

Fissure

Forlorn

Jandelar

Lag

Lucid

Masonry

Motley

Petulant

Phisteron

Roost

Rustling

Scrutinizing

Sentiment

Stave

Steed

Sullen

Redistac

Tarn

Thonstan

Thwart

Tumultuous

Veritable

Wane

Wended

Whim

Directions: Next, go through the list again and place a check mark beside any word that you left blank if you have seen it before or if it is familiar to you, even if you are not quite sure what it means.

Appendix B: Narrative Passages

Pilot Study

Directions: Please read the following passages and mark the scales. For example, if you are 85% sure you have understood the passage, mark the scale as follows:

0 _____ 100%

Passage

His appetite for the marvelous, and his powers of digesting it, were equally extraordinary; and both had been increased by his residence in this spell bound region. No tale was too gross or monstrous for his capacious swallow. It was often his delight, after his school was dismissed of an afternoon, to stretch himself on the rich bed of clover, bordering the little brook that whimpered by his schoolhouse, and there con over old Mather's direful tales, until the gathering dusk of evening made the printed page a mere mist before his eyes. Then, as he wended his way, by swamp and stream and awful woodland, to the farmhouse where he happened to be quartered, every sound of nature, at that witching hour, fluttered his excited imagination: the moan of the whip-poor-will from the hillside; the boding cry of the tree toad, that harbinger of storm; the dreary hooting of the screech owl; or the sudden rustling in the thicket, of birds frightened from their roost. The fireflies, too, which sparkled most vividly in the darkest places, now and then startled him, as one of uncommon brightness would stream across his path; and if, by chance, a huge blockhead of a beetle came winging his blundering flight against him, the poor varlet was ready to give up the ghost, with the idea that he was struck with a witch's token.

How confident are you in your understanding of the passage's overall meaning?

0 _____ 100%

How confident are you in your understanding of individual word meanings from the passage?

0 _____ 100%

Passage

Brad, who had a degree of rough chivalry in his nature, would fain have carried matters to open warfare, and have settled their pretensions to the lady, according to the mode of those most concise and simple reasoners, the knights errant of yore – by single combat; but Isaac was too conscious of the superior might of his adversary to enter the lists against him; he had overheard a boast of Brad, that he would "double the schoolmaster up, and lay him on a shelf of his own schoolhouse;" and he was too wary to give him an opportunity. There was something extremely provoking in this obstinately pacific system; it left Brad no alternative but to draw upon the funds of rustic waggery in his disposition, and to play off boorish practical jokes upon his rival. Isaac became the object of whimsical persecution to Brad, and his gang of rough riders. They harried his hitherto peaceful domains; smoked out his singing school, by stopping up the chimney; broke into the schoolhouse at night, in spite of its formidable fastenings of I and window stakes, and turned everything topsy-turvy, so that the poor schoolmaster began to think all the witches in the country held their meetings there. But what was still more annoying, Brad took all opportunities of turning him into ridicule in presence of his mistress, and had a scoundrel dog, whom he taught to whine in the most ludicrous manner, and introduced as a rival of Isaac's, to instruct her in psalmody.

How confident are you in your understanding of the passage's overall meaning?

0 _____ 100%

How confident are you in your understanding of individual word meanings from the passage?

0 _____ 100%

Passage

Ned, who had no relish for this strange midnight companion, and bethought himself of the adventure of Adam Brown with the galloping Hessian, now quickened his steed, in hopes of leaving him behind. The stranger, however, quickened his horse to an equal pace; Ned pulled up, and fell into a walk, thinking to lag behind – the other did the same. His heart began to sink within him; he endeavored to resume his psalm tune, but his parched tongue clove to the roof of his mouth, and he could not utter a stave. There was something in the moody and dogged silence of this pertinacious companion, that was mysterious and appalling. It was soon fearfully accounted for. On mounting a rising ground, which brought the figure of his fellow traveler in relief against the sky, gigantic in height, and muffled in a cloak, Ned was horror struck, on perceiving that he was headless! But his horror was still more increased, on observing, that the head, which should have rested on his shoulders, was carried before him on the pommel of the saddle! His terror rose to desperation; he rained a shower of kicks and blows upon Gunpowder, hoping by a sudden movement, to give his companion the slip – but the specter started full jump with him. Away, then, they dashed, through thick and thin; stones flying, and sparks flashing, at every bound.

How confident are you in your understanding of the passage's overall meaning?

0 _____ 100%

How confident are you in your understanding of individual word meanings from the passage?

0 _____ 100%

Passage

The portrait, I have already said, was that of a young girl. It was a mere head and shoulders, done in what is technically termed a vignette manner, much in the style of the favorite heads of Sully. The arms, the bosom, and even the ends of the radiant hair melted imperceptibly into the vague yet deep shadow which formed the background of the whole. The frame was oval, richly gilded and filigreed in Moresque. As a thing of art nothing could be more admirable than the painting itself. But it could have been neither the execution of the work, nor the immortal beauty of the countenance, which had so suddenly and so vehemently moved me. Least of all, could it have been that my fancy, shaken from its half slumber, had mistaken the head for that of a living person. I saw at once that the peculiarities of the design, of the vignetting, and of the frame, must have instantly dispelled such an idea – must have prevented even its momentary entertainment. Thinking earnestly upon these points, I remained, for an hour perhaps, half sitting, half reclining, with my vision riveted upon the portrait. At length, satisfied with the true secret of its effect, I fell back within the bed. I had found the spell of the picture in an absolute life-likeness of expression, which, at first startling, finally confounded, subdued, and appalled me.

How confident are you in your understanding of the passage's overall meaning?

0 _____ 100%

How confident are you in your understanding of individual word meanings from the passage?

0 _____ 100%

Passage

Shaking from my spirit what must have been a dream, I scanned more narrowly the real aspect of the building. Its principal feature seemed to be that of an excessive antiquity. The discoloration of ages had been great. Minute fungi overspread the whole exterior, hanging in a fine tangled web-work from the eaves. Yet all this was apart from any extraordinary dilapidation. No portion of the masonry had fallen and there appeared to be a wild inconsistency between its still perfect adaptation of parts, and the crumbling condition of the individual stones. In this there was much that reminded me of the specious totality of old wood-work, which has rotted for long years in some neglected vault, with no disturbance from the breath of external air. Beyond this indication of extensive decay, however, the fabric gave little token of instability. Perhaps the eye of a scrutinizing observer might have discovered a barely perceptible fissure, which, extending from the roof of the building in front, made its way down the wall in a zigzag direction, until it became lost in the sullen waters of the tarn.

Noticing these things, I rode over a short causeway to the house. A servant in waiting took my horse, and I entered the Gothic archway of the hall. A valet of stealthy step thence conducted me in silence through many dark and intricate passages in my progress to the studio of his master.

How confident are you in your understanding of the passage's overall meaning?

0 _____ 100%

How confident are you in your understanding of individual word meanings from the passage?

0 _____ 100%

Passage

It may seem strange that in spite of the continual anxiety occasioned me by the rivalry of Wilson, and his intolerable spirit of contradiction, I could not bring myself to hate him altogether. We had, to be sure, nearly every day a quarrel in which, yielding me publicly the palm of victory, he, in some manner, contrived to make me feel that it was he who had deserved it; yet a sense of pride on my part, and a veritable dignity on his own, kept us always upon what are called "speaking terms," while there were many points of strong congeniality in our tempers, operating to awake in me a sentiment our position alone, perhaps, prevented from ripening into friendship. It is difficult indeed, to define, or even to describe, my real feelings toward him. They formed a motley and heterogeneous admixture – some petulant animosity, which was not yet hatred, some esteem, more respect, much fear, with a world of uneasy curiosity. To the moralist it will be necessary to say, in addition, that Wilson and myself were the most inseparable of companions.

It was no doubt the anomalous state of affairs existing between us, which turned all my attacks upon him (and there were many, either open or covert) into the channel of banter or practical joke (giving pain while assuming the aspect of mere fun) rather than into a more serious and determined hostility.

How confident are you in your understanding of the passage's overall meaning?

0 _____ 100%

How confident are you in your understanding of individual word meanings from the passage?

0 _____ 100%

Dissertation Study

Directions: Please read the following passages and mark the scales. For example, if you are 85% sure you have understood the passage, mark the scale as follows:

0 _____ 100%

Passage

His appetite for the marvelous, and his powers of digesting it, were equally extraordinary; and both had been increased by his residence in this spell bound region. No tale was too gross or monstrous for his capacious swallow. It was often his delight, after his school was dismissed of an afternoon, to stretch himself on the rich bed of clover, bordering the little brook that whimpered by his schoolhouse, and there con over old Mather's direful tales, until the gathering dusk of evening made the printed page a mere mist before his eyes. Then, as he wended his way, by swamp and stream and awful woodland, to the farmhouse where he happened to be quartered, every sound of nature, at that witching hour, fluttered his excited imagination: the moan of the whip-poor-will from the hillside; the boding cry of the tree toad, that harbinger of storm; the dreary hooting of the screech owl; or the sudden rustling in the thicket, of birds frightened from their roost. The fireflies, too, which sparkled most vividly in the darkest places, now and then startled him, as one of uncommon brightness would stream across his path; and if, by chance, a huge blockhead of a beetle came winging his blundering flight against him, the poor varlet was ready to give up the ghost, with the idea that he was struck with a witch's token.

How confident are you in your understanding of the passage's overall meaning?

0 _____ 100%

How confident are you in your understanding of individual word meanings from the passage?

0 _____ 100%

Summarize the main idea of the passage in one or two sentences.

Passage

Ned, who had no relish for this strange midnight companion, and bethought himself of the adventure of Adam Brown with the galloping Hessian, now quickened his steed, in hopes of leaving him behind. The stranger, however, quickened his horse to an equal pace; Ned pulled up, and fell into a walk, thinking to lag behind – the other did the same. His heart began to sink within him; he endeavored to resume his psalm tune, but his parched tongue clove to the roof of his mouth, and he could not utter a stave. There was something in the moody and dogged silence of this pertinacious companion, that was mysterious and appalling. It was soon fearfully accounted for. On mounting a rising ground, which brought the figure of his fellow traveler in relief against the sky, gigantic in height, and muffled in a cloak, Ned was horror struck, on perceiving that he was headless! But his horror was still more increased, on observing, that the head, which should have rested on his shoulders, was carried before him on the pommel of the saddle! His terror rose to desperation; he rained a shower of kicks and blows upon Gunpowder, hoping by a sudden movement, to give his companion the slip – but the specter started full jump with him. Away, then, they dashed, through thick and thin; stones flying, and sparks flashing, at every bound.

How confident are you in your understanding of the passage's overall meaning?

0 _____ 100%

How confident are you in your understanding of individual word meanings from the passage?

0 _____ 100%

Summarize the main idea of the passage in one or two sentences.

Passage

Shaking from my spirit what must have been a dream, I scanned more narrowly the real aspect of the building. Its principal feature seemed to be that of an excessive antiquity. The discoloration of ages had been great. Minute fungi overspread the whole exterior, hanging in a fine tangled web-work from the eaves. Yet all this was apart from any extraordinary dilapidation. No portion of the masonry had fallen and there appeared to be a wild inconsistency between its still perfect adaptation of parts, and the crumbling condition of the individual stones. In this there was much that reminded me of the specious totality of old wood-work, which has rotted for long years in some neglected vault, with no disturbance from the breath of external air. Beyond this indication of extensive decay, however, the fabric gave little token of instability. Perhaps the eye of a scrutinizing observer might have discovered a barely perceptible fissure, which, extending from the roof of the building in front, made its way down the wall in a zigzag direction, until it became lost in the sullen waters of the tarn.

Noticing these things, I rode over a short causeway to the house. A servant in waiting took my horse, and I entered the Gothic archway of the hall. A valet of stealthy step thence conducted me in silence through many dark and intricate passages in my progress to the studio of his master.

How confident are you in your understanding of the passage's overall meaning?

0 _____ 100%

How confident are you in your understanding of individual word meanings from the passage?

0 _____ 100%

Summarize the main idea of the passage in one or two sentences.

Passage

It may seem strange that in spite of the continual anxiety occasioned me by the rivalry of Wilson, and his intolerable spirit of contradiction, I could not bring myself to hate him altogether. We had, to be sure, nearly every day a quarrel in which, yielding me publicly the palm of victory, he, in some manner, contrived to make me feel that it was he who had deserved it; yet a sense of pride on my part, and a veritable dignity on his own, kept us always upon what are called "speaking terms," while there were many points of strong congeniality in our tempers, operating to awake in me a sentiment our position alone, perhaps, prevented from ripening into friendship. It is difficult indeed, to define, or even to describe, my real feelings toward him. They formed a motley and heterogeneous admixture – some petulant animosity, which was not yet hatred, some esteem, more respect, much fear, with a world of uneasy curiosity. To the moralist it will be necessary to say, in addition, that Wilson and myself were the most inseparable of companions.

It was no doubt the anomalous state of affairs existing between us, which turned all my attacks upon him (and there were many, either open or covert) into the channel of banter or practical joke (giving pain while assuming the aspect of mere fun) rather than into a more serious and determined hostility.

How confident are you in your understanding of the passage's overall meaning?

0 _____ 100%

How confident are you in your understanding of individual word meanings from the passage?

0 _____ 100%

Summarize the main idea of the passage in one or two sentences.

Appendix C: Word Knowledge Posttest
Pilot Study

Directions: Write a definition or short description for each word. Please make your definitions as clear as possible so that I know you understand the meaning of the word. If you are unsure of a word's meaning, write your best guess.

Boding

How confident are you in the accuracy of your response?

0 _____ 100%

Boorish

0 _____ 100%

Capacious

0 _____ 100%

Clove

0 _____ 100%

Congeniality

0 _____ 100%

Contrived

0 _____ 100%

Countenance

0 _____ 100%

Dilapidation

0 _____ 100%

Dispelled

0 _____ 100%

Filigreed

0 _____ 100%

Fissure

0 _____ 100%

Gilded

0 _____ 100%

Motley

0 _____ 100%

Pacific

0 _____ 100%

Pertinacious

0 _____ 100%

Petulant

0 _____ 100%

Pommel

0 _____ 100%

Psalmody

0 _____ 100%

Specious

0 _____ 100%

Specter

0 _____ 100%

Stave

0 _____ 100%

Sullen

0 _____ 100%

Tarn

0 _____ 100%

Varlet

0 _____ 100%

Vehemently

0 _____ 100%

Veritable

0 _____ 100%

Vignette

0 _____ 100%

Waggery

0 _____ 100%

Wended

0 _____ 100%

Withe

0 _____ 100%

Dissertation Study

Directions: Write a definition or short description for each word. Please make your definitions as clear as possible so that I know you understand the meaning of the word. If you are unsure of a word's meaning, write your best guess.

Once you have written your best guess for each word, mark each scale to indicate how confident you are in the accuracy of your response.

Animosity

0 _____ 100%

Anomalous

0 _____ 100%

Banter

0 _____ 100%

Benfactor

0 _____ 100%

Blundering

0 _____ 100%

Clove

0 _____ 100%

Con

0 _____ 100%

Dilapidation

0 _____ 100%

Dispelled

0 _____ 100%

Endeavored

0 _____ 100%

Fissure

0 _____ 100%

Lag

0 _____ 100%

Lucid

0 _____ 100%

Masonry

0 _____ 100%

Motley

0 _____ 100%

Roost

0 _____ 100%

Rustling

0 _____ 100%

Scrutinizing

0 _____ 100%

Sentiment

0 _____ 100%

Stave

0 _____ 100%

Steed

0 _____ 100%

Tarn

0 _____ 100%

Tumultuous

0 _____ 100%

Wended

0 _____ 100%

Whim

0 _____ 100%

Appendix D: Think Aloud Protocol

Many people talk to themselves while they read. What we are interested in for this study is what you think and do while you read a text. You can decide for yourself whether you would like to read the text silently or out loud, or do something of both. Do whatever feels most natural for you. We are only interested in what you are thinking and doing as you read the text. For example, if you chose to reread parts of the text, please say so. If something in the text reminds you of something you already know or prior experiences, please say so. If something is confusing, please let us know that, too. If you are quiet for a period of time, I will remind you to say what you are thinking. Do you have any questions?

To get used to thinking aloud, we have a short practice passage for you. We will not record this one and you can take your time and get used to how it feels. Now, what I would like you to do is read the passage and say out loud what you are thinking and doing.

Additional instructions for the word knowledge posttest:

Please continue to say what you are thinking while you generate definitions and make confidence judgments about your answers. You can go back to the passages if that is helpful. Please take your best guess at definitions or descriptions for the words on this list and say out loud what you are thinking and doing.

Appendix E: Practice Passage

Passage

Brad, who had a degree of rough chivalry in his nature, would fain have carried matters to open warfare, and have settled their pretensions to the lady, according to the mode of those most concise and simple reasoners, the knights errant of yore – by single combat; but Isaac was too conscious of the superior might of his adversary to enter the lists against him; he had overheard a boast of Brad, that he would "double the schoolmaster up, and lay him on a shelf of his own schoolhouse;" and he was too wary to give him an opportunity. There was something extremely provoking in this obstinately pacific system; it left Brad no alternative but to draw upon the funds of rustic waggery in his disposition, and to play off boorish practical jokes upon his rival. Isaac became the object of whimsical persecution to Brad, and his gang of rough riders. They harried his hitherto peaceful domains; smoked out his singing school, by stopping up the chimney; broke into the schoolhouse at night, in spite of its formidable fastenings of withe and window stakes, and turned everything topsy-turvy, so that the poor schoolmaster began to think all the witches in the country held their meetings there. But what was still more annoying, Brad took all opportunities of turning him into ridicule in presence of his mistress, and had a scoundrel dog, whom he taught to whine in the most ludicrous manner, and introduced as a rival of Isaac's, to instruct her in psalmody.

How confident are you in your understanding of the passage's overall meaning?

0 _____ 100%

How confident are you in your understanding of individual word meanings from the passage?

0 _____ 100%

Summarize the main idea of the passage in one or two sentences.

Appendix F: Coding Scheme for the Verbal Protocol Analysis

Strategic behaviors

- Reading aloud
- Re-reading
- Adjusting reading rate when re-reading – speeding up or slowing down
- Skimming (reading aloud while skipping portions)
- Guessing the meaning of a word in context [“Erroneous I think means things that are not necessarily factual.”]
- Predicting [“Okay, now it’s going to summarize that.”]
- Questioning [“What would happen if you do it either direction?”]
- Arguing with text [“it also, it really depends on your knowledge of the subject, ‘cause if you don’t know much about it, it won’t seem vague or improbable evidence.”]
- Underlining or other marking on the text [“Underlining intuition and authority.”]
- Using text feature [“I’m looking for it in the table.”]
- Rehearsing (repeating information to maintain it in memory) [“So that’s type one error. Type two. Okay. Type one, type two.”]
- Restating (paraphrase) or repeating text information
 - local (word, phrase, sentence level) [“So significance level can increase or decrease the type one error.”]
 - global (paragraph, passage level) [“So basically it introduces about, um, how the scientific approach differs from just, uh, intuition and authority.”]

- Making connections
 - to background knowledge [“We learned about peer review in class, it’s when other people kind of look at your results and confirm it.”]
 - to personal experience [“That happened to my sister.”]
 - to prior text [“Intuition and authority are the things I just read about.”]
 - to topic knowledge test [“This probably relates to scientific skepticism, which was the thing, a question on one of those tests I just took.”]
 - to research task [“I’m not gonna have much to write about this.”]
- Interpreting (a statement requiring reasoning beyond information in the text to build text meaning) [“So, that’s just talking about the confidence interval.”]
- Elaborating (a statement requiring the use of additional information not explicitly in the text to build beyond text meaning or pursue a non-text related train of thought) [“what if, um, what if that one person just, like, stole a biscuit or something.”]

Monitoring/Evaluative behaviors

- Evaluating comprehension (positive or negative) [“I’m already confused by this passage.”]
- Evaluating agreement with text (positive or negative) [“That’s definitely true.”]
- Evaluating text quality [“That’s a good way to describe it.”]
- Evaluation of interest (positive or negative) [“The first part was kind of interesting.”]

- Evaluation of importance of text [“I feel like that’s important, with the, to know for later on this semester.”]
- Evaluation of task difficulty [“In order for me to really realize what is going on here, I would have to sit down and study this stuff.”]
- Monitoring task completion status [“Okay, I’m done.”]

Other

- Expression of empathy (sympathy or feelings felt or imputed to others) [“That’s, that’s really nice when people adopt children.”]
- Expression of amusement [“Um, [laughs] I was thinking it was funny.”]
- Expression of surprise [“Surprised by the findings.”]
- No code (not enough information available to determine a code, as when comments are partially unintelligible or fragmentary) [... Okay, so that’s [unintelligible] intuition and Aristotle.”]

Appendix G: Demographics Sheet

This information is for the sole purpose of reporting overall sample characteristics, and will not be used to identify or categorize participants in any way.

Gender: ___ Male ___ Female

Age: _____

College Year:

___ Freshman

___ Sophomore

___ Junior

___ Senior

___ Other (please specify): _____

Major: _____ **Overall GPA:** _____

Ethnicity (check all that apply):

___ African American

___ American Indian

___ Asian/Pacific Islander American

___ European American

___ Hispanic American

___ Other (please specify): _____

Are you a native English speaker?

___ Yes

___ No

Appendix H: Word Characteristics from WordNet 2.1

Word	Function	Familiarity	Polysemy count
Admonish (v)	Filler	Uncommon	3
Animosity (n)	Target	Very rare	1
Anomalous (adj)	Target	Very rare	1
Banter (n)	Target	Very rare	1
Banter (v)	Target	Very rare	1
Benefactor (n)	Target	Very rare	1
Blundering (v)	Target	Uncommon	3
Capacious (adj)	Filler	Very rare	1
Clove (n)	Target	Uncommon	4
Clove (cleave) (v)	Target	Uncommon	3
Con (n)	Target	Uncommon	3
Con (v)	Target	Rare	2
Con (adv)	Target	Very rare	1
Congeniality (n)	Filler	Rare	2
Devernal	Pseudoword		
Dilapidation (n)	Target	Rare	2
Dispelled (v)	Filler	Rare	2
Edarthic	Pseudoword		
Endeavored (v)	Target	Very rare	1
Fissure (n)	Target	Uncommon	3

Word	Function	Familiarity	Polysemy count
Fissure (v)	Target	Very rare	1
Forlorn (adj)	Filler	Very rare	1
Jandelar	Pseudoword		
Lag (n)	Target	Uncommon	3
Lag (v)	Target	Uncommon	4
Lucid (adj)	Target	Uncommon	4
Masonry (n)	Target	Uncommon	3
Motley (n)	Target	Uncommon	3
Motley (v)	Target	Rare	2
Motley (adj)	Target	Rare	2
Petulant (adj)	Filler	Very rare	1
Phisteron	Pseudoword		
Roost (n)	Target	Rare	2
Roost (v)	Target	Rare	2
Rustling (n)	Target	Rare	2
Rustling (v)	Target	Rare	2
Rustling (adj)	Target	Very rare	1
Scrutinizing (v)	Target	Rare	2
Sentiment (n)	Target	Rare	2
Stave (n)	Target	Uncommon	3
Stave (v)	Target	Rare	2
Steed (n)	Target	Very rare	1

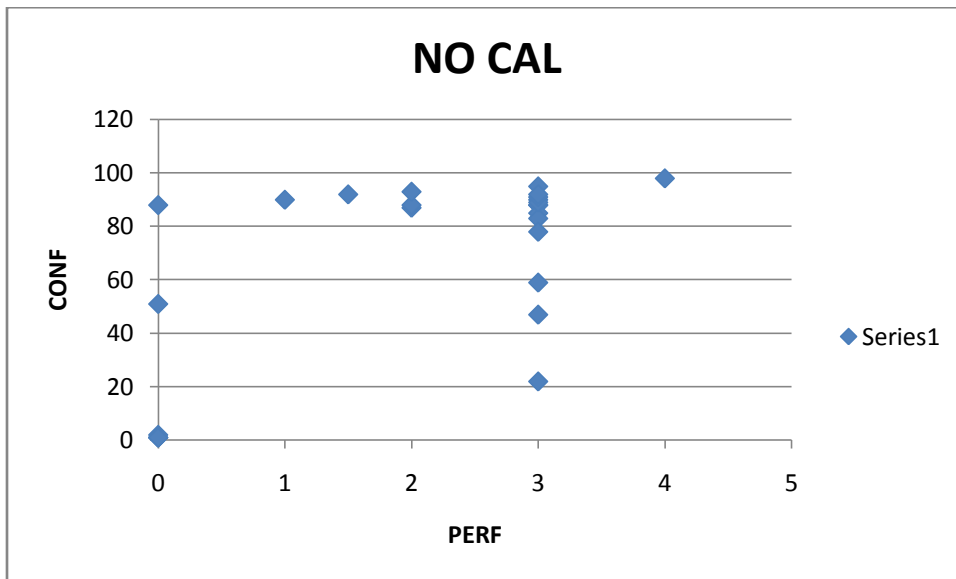
Word	Function	Familiarity	Polysemy count
Sullen (adj)	Filler	Rare	2
Redistac	Pseudoword		
Tarn (n)	Target	Very rare	1
Thonstan	Pseudoword		
Thwart (n)	Filler	Very rare	1
Thwart (v)	Filler	Very rare	1
Tumultuous (adj)	Filler	Very rare	1
Veritable (adj)	Filler	Rare	2
Wane (n)	Filler	Very rare	1
Wane (v)	Filler	Uncommon	3
Wended (v)	Target	Very rare	1
Whim (n)	Target	Rare	2

Appendix I: Coding Scheme for Main Idea Response Ratings

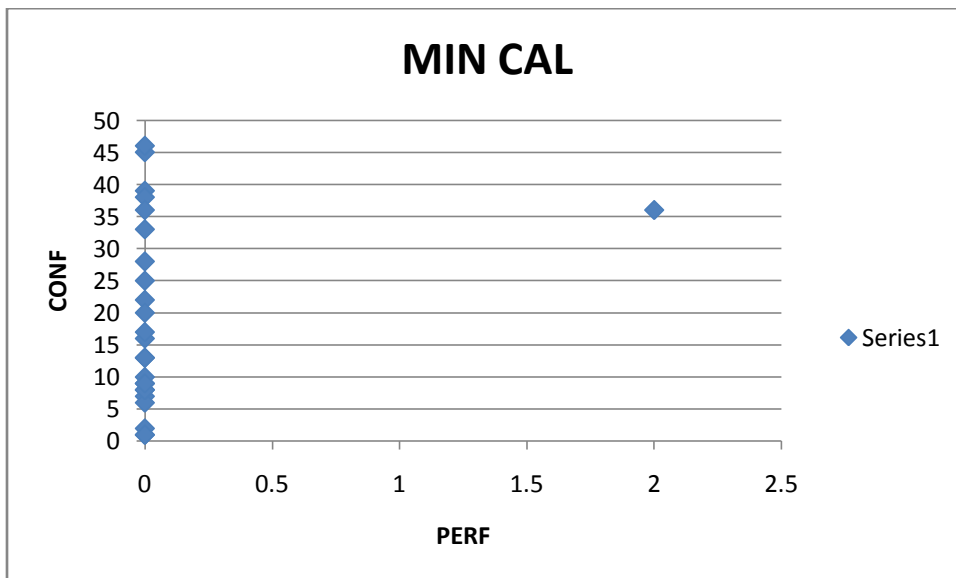
Rating	Description
0	No correct features of the main idea
.5	Correct and incorrect features of the main idea
1	All or most correct features of the main idea (no incorrect features)

Appendix J: Calibration Scatterplots

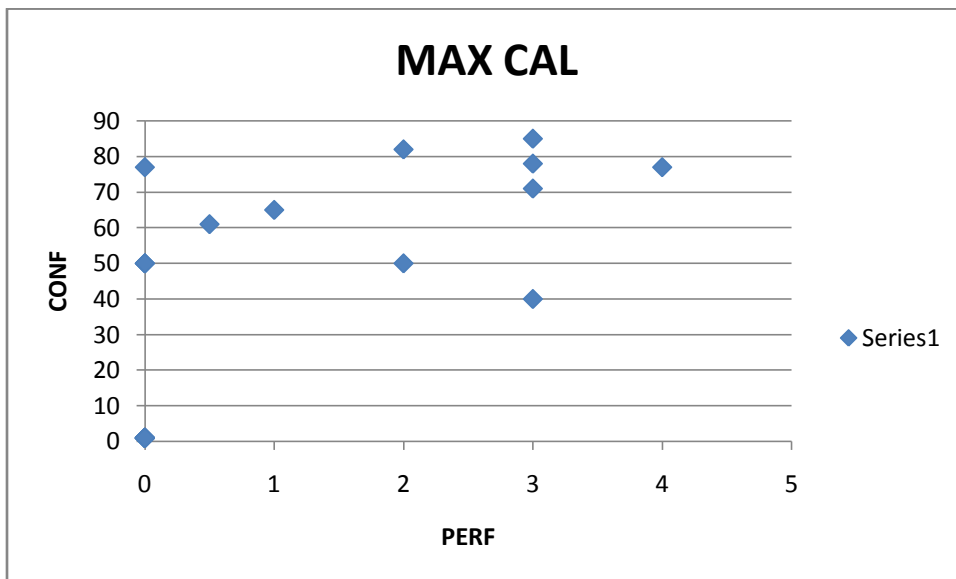
Participant with a Spearman's rho = .08 (No Calibration)



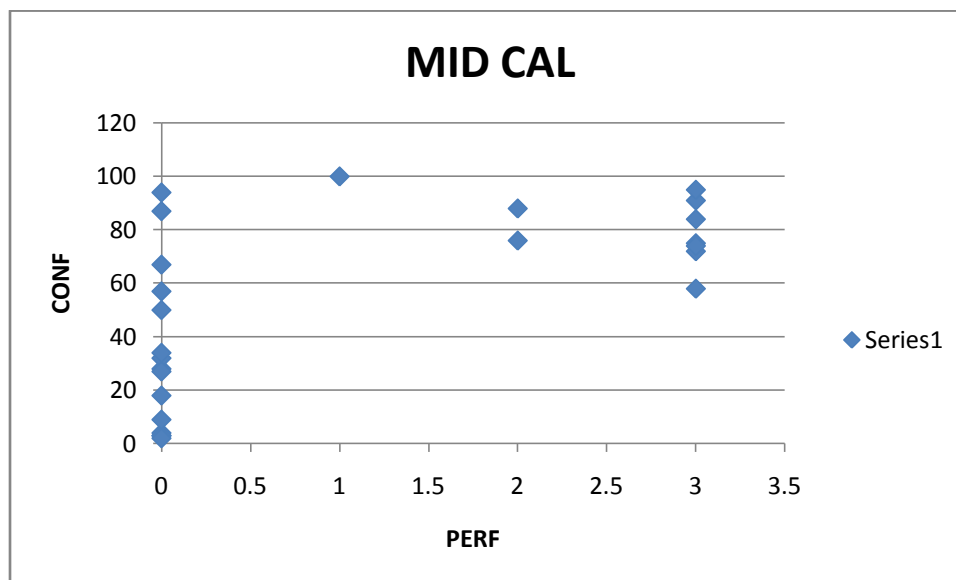
Participant with Spearman's rho = -.61 (Very Poor Calibration)



Participant with Spearman's rho = .84 (Very High Calibration)



Participant with Spearman's rho = .38 (Mid-level of Calibration)



REFERENCES

- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: The MIT Press.
- Albaum, G., Best, R., & Hawkins, D. I. (1981). Continuous vs. discrete semantic differential rating scales. *Psychological Reports, 49*, 83-86.
- Alexander, P. A. (2004). A model of domain learning: Reinterpreting expertise as a multidimensional, multistage process. In D. Y. Dai & R. J. Sternberg (Eds.), *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development* (pp. 273-298). Mahwah, NJ: Lawrence Erlbaum Associates.
- Alexander, P. A. (2005). The path to competence: A lifespan developmental perspective on reading. *Journal of Literacy Research, 37*, 413-436.
- Alexander, P. A., Murphy, P. K., Woods, B. S., Duhon, K. E., & Parker, D. (1997). College instruction and concomitant changes in students' knowledge, interest, and strategy use: A study of domain learning. *Contemporary Educational Psychology, 22*, 125-146.
- Alexander, P. A., & Jetton, T. L. (1996). The role of importance and interest in the processing of text. *Educational Psychology Review, 8*, 89-121.
- Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage Publications.
- Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77-117). Newark, DE: International Reading Association.

- Anderson, R. C., & Nagy, W. E. (1991). Word meanings. In R. Barr, M. L. Kamil, & P. Mosenthal, & P. D. Pearson (Eds.). *Handbook of reading research* (Vol. 2, pp. 690-724). White Plains, NY: Longman.
- Baker, L. (1984). Spontaneous versus instructed use of multiple standards for evaluating comprehension: Effects of age, reading proficiency, and type of standard. *Journal of Experimental Child Psychology*, 38, 289-311.
- Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. D. Pearson, M. Kamil, R. Barr, & P. Mosenthal (Eds.), *Handbook of reading research* (Vol. 1, pp. 353-394). White Plains, NY: Longman.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Baumann, J. F., Edwards, E. C., Boland, E. M., Olejnik, S., & Kameenui, E. J. (2003). Vocabulary tricks: Effects of instruction in morphology and context on fifth-grade students' ability to derive and infer word meanings. *American Educational Research Journal*, 40, 447-494.
- Beck, I. L., McKeown, M. G., & McCaslin, E. (1983). All contexts are not created equal. *Elementary School Journal*, 83, 177-181.
- Bentler, P. M. (1998). *EQS: A structural equations program*. Encino, CA: Multivariate Software, Inc.
- Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, 45, 122-159.

- Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, 15, 1-40.
- Brown, R. W. (1957). Linguistic determinism and the part of speech. *Journal of Abnormal and Social Psychology*, 55, 1-5.
- Brown, J. I., Fishco, V. V., & Hanna, G. S. (1993). *Nelson-Denny reading test: Manual for scoring and interpretation, Forms G & H*. Itasca, IL: Riverside Publishing.
- Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage word frequency book*. Boston: Houghton Mifflin Company.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407-428.
- Curtis, M. E. (2006). The role of vocabulary instruction in adult basic education. In J. Comings, B. Garner, & C. Smith (Eds.), *Review of adult learning and literacy: Connecting research, policy, and practice* (Vol. 6, pp. 43-69). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Davis, F. B. (1944). Fundamental factors of comprehension in reading. *Psychometrika*, 9, 185-197.

- Dinsmore, D. L., & Parkinson, M. M. (2011). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. In P. Alexander (Chair), *Calibrating calibration: Conceptualization, measurement, calculation, and context*. Symposium presented at the annual meeting of the American Educational Research Association, New Orleans.
- Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second-order judgments about judgments of learning. *The Journal of General Psychology*, *132*, 335-346.
- Dunlosky, J., Rawson, K. A., Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language*, *52*, 551-565.
- Durkin, P. M. (1990). Vocabulary acquisition in context reconsidered: The effect of word type and exposure level on the learning of unknown words. *Dissertation Abstracts International*, *51*(08A). (University Microfilms No. AAI9102613).
- Durso, F. T., & Coggins, K. A. (1991). Organized instruction for the improvement of word knowledge skills. *Journal of Educational Psychology*, *83*, 108-112.
- Durso, F. T., & Shore, W. J. (1991). Partial knowledge of word meanings. *Journal of Experimental Psychology: General*, *120*, 190-202.
- Enders, C. K. (2006). Analyzing structural equation models with missing data. In G. R. Hancock and R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 315-344). Greenwich, CT: Information Age Publishing.

- Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: The MIT Press.
- Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation models. In G. R. Hancock & R. O. Mueller (Eds.). *A second course in structural equation modeling* (pp. 269 – 314). Greenwich, CT: Information Age.
- Fischer, U. (1994). Learning words from context and dictionaries: An experimental comparison. *Applied Psycholinguistics*, *15*, 551-574.
- Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 552-564.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*, 906-911.
- Fukkink, R. G. (2005). Deriving word meaning from written context: a process analysis. *Learning and Instruction*, *15*, 23-43.
- Gardner, D. (2004). Vocabulary input through extensive reading: A comparison of words found in children's narrative and expository reading materials. *Applied Linguistics*, *25*, 1-37.
- Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 702-718.
- Glenberg, A. M., Sanocki, T., & Epstein, W. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General*, *116*, 119-136.

- Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics, 11*, 341-363.
- Graves, M. (1984). Selecting vocabulary to teach in the intermediate and secondary grades in J. Flood (Ed.), *Promoting reading comprehension* (pp. 245-260). Newark, DE: International Reading Association.
- Graves, M. (2000). A vocabulary program to complement and bolster a middle-grade comprehension program. In B. Taylor, M. Graves, & P. van den Broek (Eds.), *Reading for meaning: Fostering comprehension in the middle grades* (pp. 116-135). Newark, DE: International Reading Association.
- Graves, M. (2006). *The vocabulary book: Learning and instruction*. New York: Teachers College Press, IRA, and NCTE.
- Hidi, S. (2001). Interest, reading, and learning: Theoretical and practical considerations. *Educational Psychology Review, 13*, 191-209.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1-55.
- Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning, 4*, 161-176.
- Irving, W. (1978). *The complete works of Washington Irving: The sketch book of Geoffery Crayon, Gent.* (H. Springer, Ed.) Boston: Twayne Publishers.
- Jenkins, J.R., & Dixon, R. (1983). Vocabulary learning. *Contemporary Educational Psychology, 8*, 237-260.

- Jenkins, J. R., Stein, M. L., & Wysocki, K. (1984). Learning vocabulary through reading. *American Educational Research Journal*, 21, 767-787.
- Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.
- Konopak, B., Sheard, C., Longman, D., Lyman, B., Slaton, E., & Atkinson, R. (1987). Incidental versus intentional word learning from context. *Reading Psychology*, 8, 7-21.
- Lennon, C., & Burdick, H. (2004). *The lexile framework as an approach for reading measurement and success*. Durham, NC: MetaMetrics, Inc.
- Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20, 159-183.
- Lichtenstein, S. & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, 26, 149-171.
- Lin, L., Moore, D., & Zabucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology*, 22, 111-128.
- Lin, L., & Zabucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, 23, 345-391.

- Lubliner, S., & Smetana, L. (2005). The effects of comprehensive vocabulary instruction on Title I students' metacognitive word-learning skills and reading comprehension. *Journal of Literacy Research, 37*, 163-200.
- Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 663-679.
- Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 4*, 723-731.
- Marzano, R. J. (2004). The developing vision of vocabulary instruction. In J. F. Baumann & E. J. Kame'enui (Eds.), *Vocabulary instruction: Research to practice* (pp. 100-117). New York: The Guilford Press.
- McCrudden, M. T., Magliano, J. P., & Schraw, G. (2010). Exploring how relevance instructions affect personal reading intentions, reading goals and text processing: A mixed methods study. *Contemporary Educational Psychology, 35*, 229-241.
- McKeown, M. G. (1985). The acquisition of word meaning from context by children of high and low ability. *Reading Research Quarterly, 20*, 482-496.
- McKeown, M. G., Beck, I. L., Omanson, R. C., & Perfetti, C. (1983). The effects of long-term vocabulary instruction on reading comprehension: A replication. *Journal of Reading Behavior, 15*, 3-18.
- McNamara, D.S., Louwerse, M.M., Cai, Z., & Graesser, A. (2005, January 1). Coh-Metrix version 1.4. Retrieved [May 16, 2011], from <http://cohmetrix.memphis.edu>.

- Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38, 39-41.
- Nagy, W. E., & Anderson, R. C. (1984). The number of words in printed school English. *Reading Research Quarterly*, 19, 304-330.
- Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, 24, 237-270.
- Nagy, W. E., Herman, P. A., & Anderson, R.C. (1985). Learning words from context. *Reading Research Quarterly*, 20, 233-253.
- Nagy, W. E., & Herman, P. A. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 19-35). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Nagy, W. E., & Scott, J. (2000). Vocabulary processing. In M. Kamil, P. Mosenthal, P. D. Pearson & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 269-284). Mahwah, NJ: Lawrence Erlbaum Associates.
- National Reading Panel. (2000). *National reading panel: Reports of the subgroups*. NIH Publication.
- Nietfeld, J. L., & Schraw, J. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research*, 95, 131-142.

- Paris, S. G., Wasik, B. A., & Turner, J. C. (1996). The development of strategic readers. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol 2., pp. 609-640). Hillsdale, NJ: Lawrence Erlbaum Associate, Inc.
- Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451-502). San Diego, CA: Academic Press.
- Pintrich, P. R., & Zusho, A. (2002) The development of academic self-regulation: The role of cognitive and motivational factors. In A. Wigfield, J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 249-284). San Diego, CA: Academic Press.
- Poe, E. A. (2004). *The tales of Edgar Allan Poe*. New York: Simon & Schuster.
- Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- RAND Reading Study Group. (2002). *Reading for understanding: Toward and R&D program in reading comprehension*. Santa Monica, CA: RAND Education.
- Ronis, D. L. & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, 40, 193-218.
- Rowling, J. K. (2000). *Harry Potter and the goblet of fire*. New York: Scholastic, Inc.
- Schraw, G., & Bruning, R. (1999). How implicit models of reading affect motivation to read and reading engagement. *Scientific Studies of Reading*, 3, 281-302.

- Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*, 351-371.
- Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology, 18*, 455-463.
- Schwanenflugel, P. J., Stahl, S. A., & McFalls, E. L. (1997). Partial word knowledge and vocabulary growth during reading comprehension. *National Reading Research Center Universities of Georgia and Maryland*, Reading research report no. 76.
- Shore, W. J., & Durso, F. T. (1990). Partial knowledge in vocabulary acquisition: General constraints and specific detail. *Journal of Educational Psychology, 82*, 315-318.
- Stahl, S.A. (1989). Task variations and prior knowledge in learning word meanings from context. In S. McCormick, & J. E. Zutell (Eds.), *Cognitive and social perspectives for literacy research and instruction* (pp. 197-204). Chicago: National Reading Conference.
- Stahl, S. A. (1991). Beyond the instrumental hypothesis: Some relationships between word meaning and comprehension. In P. J. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 157-186). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Stahl, S. A. (1999). *Vocabulary development*. Cambridge, MA: Brookline Books.
- Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*, 72-110.

- Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-407.
- Sternberg, R. J., & Powell, J. S. (1983). Comprehending verbal comprehension. *American Psychologist, 38*, 878-893.
- Swanborn, M. S. L., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research, 69*, 261-285.
- Swanborn, M. S. L., & de Glopper, K. (2002). Impact of reading purpose on incidental word learning from context. *Language Learning, 52*, 95-117.
- Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1267-1280.
- Van Daalen-Kapteijns, M. M., & Elshout-Mohr, M. (1981). The acquisition of word meanings as a cognitive learning process. *Journal of Verbal Learning and Verbal Behavior, 20*, 386-399.
- Van Daalen-Kapteijns, M. M., Elshout-Mohr, M., & de Glopper, K. (2001). Deriving the meaning of unknown words from multiple contexts. *Language Learning, 51*, 145-181.
- Veenman, M. V. J., Elshout, J. J., & Meijer, J. (1997). The generality vs domain-specificity of metacognitive skills in novice learning across domains. *Learning and Instruction, 7*, 187-209.

- Vosniadou, S., Pearson, P. D., & Rogers, T. (1988). What causes children's failures to detect inconsistencies in text? Representation versus comparison difficulties. *Journal of Educational Psychology, 80*, 27-39.
- Waern, Y. (1988). Thoughts on text in context: Applying the think-aloud method to text processing. *Text, 8*, 327-350.
- Weaver, C. A. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 214-222.
- Yates, F. J., Zhu, Y., Ronis, D. L., Wang, D., Shinotsuka, H., & Toda, M. (1989). Probability judgment accuracy: China, Japan, and the United States. *Organizational Behavior and Human Decision Processes, 43*, 145-171.
- Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*, 3-17.