9-12-2006

The Use of Item Response Theory to Assess Adults' Postdiction Accuracy

Andrea Mueller Cummings

ACCEPTANCE

This dissertation, THE USE OF ITEM RESPONSE THEORY TO ASSESS ADULTS' POSTDICTION ACCURACY, by ANDREA MUELLER CUMMINGS, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chair, as representatives of the faculty, certify that this dissertation has met all the standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

_____
Karen M. Zabrucky, Ph.D.
Committee Chair

_____
Laura D. Fredrick, Ph.D.
Committee Member

_____
John H. Neel, Ph.D.
Committee Member

_____
Dennis N. Thompson, Ph.D.
Committee Member

_____
Date

_____
Peggy A. Gallagher, Ph.D.
Chair, Department of Educational Psychology
and Special Education

_____
Ronald P. Colarusso, Ed.D.
Dean

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Andrea Mueller Cummings
1016 Sandy Lane Drive
Alpharetta, GA 30022

The director of this dissertation is:

Karen M. Zabrucky
Department of Educational Psychology and Special Education
College of Education
Georgia State University
Atlanta, GA 30303-3083

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education's director of graduate studies and research, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____Andrea Mueller Cummings_____
Signature of Author

VITA

Andrea Mueller Cummings

ADDRESS: 1016 Sandy Lane Drive
Alpharetta, GA 30022

EDUCATION:

Ph.D.	2006	Georgia State University Educational Psychology
M.S.	1997	Georgia State University Educational Psychology
B.S.	1994	Georgia State University Psychology

PROFESSIONAL EXPERIENCE:

2000-Present	Graduate Teaching Assistant, Georgia State University
1999-2003	Graduate Research Assistant, Georgia State University

SELECTED PRESENTATIONS AND PUBLICATIONS:

Moore, D., Zabrucky, K. M., Cummings, A. M., & Agler, L. *Comprehension performance: Relations with self-reported evaluation and regulation strategies.* Poster presented at 2004 Annual Meeting of Southeastern Psychological Association, Atlanta, GA.

Zabrucky, K. M., & Cummings, A. M. (2004). Metacognition, In C. Fisher & R. Lerner (Eds.), *Encyclopedia of applied developmental science* (pp. 723-727). Thousand Oaks, CA: Sage.

Moore, D., Zabrucky, K. M., Cummings, A. M., & Lin, L. *Relations between self-assessed performance and objective comprehension performance.* Poster presented at 2003 Annual Meeting of Southeastern Psychological Association, New Orleans, LA.

Moore, D., Zabrucky, K. M., & Cummings, A. M. *The revised metacomprehension scale: Validity evidence.* Poster presented at 2002 Annual Meeting of Southeastern Psychological Association, Orlando, FL.

Moore, D., Zabrucky, K. M., & Cummings, A. M. *Revision of the metacomprehension scale: Reliability and factorial validity.* Poster presented at 2000 Annual Meeting of Southeastern Psychological Association, New Orleans, LA.

Cummings, A. M. *Children's suggestibility: A meta-analysis.* Poster presented at 1999 Annual Meeting of Southeastern Psychological Association, Savannah, GA.

ABSTRACT

THE USE OF ITEM RESPONSE THEORY TO ASSESS
ADULTS' POSTDICTION ACCURACY
by
Andrea Mueller Cummings

Researchers interested in metacognition of text comprehension (*metacomprehension*) have investigated both a knowledge and a monitoring component. Knowledge of comprehension consists of one's awareness of person, strategy, and task variables and is investigated primarily through interviews and questionnaires. Monitoring of comprehension consists of two equally important abilities: evaluation and regulation. Evaluation involves adults' ability to assess their understanding during reading, whereas regulation involves their ability to use compensatory strategies to resolve comprehension failures. Monitoring of comprehension is assessed through a variety of paradigms, such as on-line performance measures, error detection, and calibration.

Researchers interested in adults' evaluation ability have frequently employed a calibration paradigm, in which adults are asked to take a comprehension test after reading one or more passages and make confidence judgments about their future test performance (*predictions*) or past test performance (*postdictions*). Findings indicate that adults are generally poor at evaluating their comprehension, and that a number of variables may influence their performance. However, findings have often been inconsistent, and a clearer picture of adults' ability is needed.

Item Response Theory (*IRT*) is a modern psychometric approach that has been successfully applied in psychological and educational research. An IRT-based comprehension test may provide a better measure of comprehension than those used in prior research. The main purpose of this study was to develop an IRT-based comprehension test for use in calibration studies. Students were also asked to report their guessing behavior, which was analyzed to determine if guessing influenced postdiction accuracy.

Undergraduate and graduate students (n=1006) completed a comprehension test, made postdictions after each item, and reported their guessing behavior. Calibration accuracy was measured by comparing students' test scores and postdictions. Factor analysis and a scree test were used to determine unidimensionality of the data, and chi square statistics were used to determine item fit. The comprehension test was found to be appropriate for distinguishing students at the low end of the ability continuum, but additional items need to be developed to discriminate among students at higher ability levels. Guessing scores were modestly, but significantly ($p < .01$) correlated with both comprehension performance and postdiction accuracy.

THE USE OF ITEM RESPONSE THEORY TO ASSESS ADULTS' POSTDICTION ACCURACY

by
Andrea Mueller Cummings

A Dissertation

Presented in Partial Fulfillment of Requirements for the
Degree of
Doctor of Philosophy
in
Educational Psychology
in
the Department of Educational Psychology and Special Education
in
the College of Education
Georgia State University

Atlanta, Georgia
2006

ACKNOWLEDGEMENTS

This dissertation is dedicated to my mother, Ingeborg Moore, and my dear friend, Marilyn Sherman, whose continued love and support inspired me to pursue my academic dreams.

Additionally, I would like to express sincere thanks to my committee chairperson, Dr. Karen Zabrucky, as well as my committee members (Dr. John Neel, Dr. Laura Fredrick, and Dr. Dennis Thompson), for their patience, guidance and mentoring through the entire process.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

CTT	Classical Test Theory
IRT	Item Response Theory
1PL	1 parameter logistic model
2PL	2 parameter logistic model
3PL	3 parameter logistic model

CHAPTER 1

AN OVERVIEW OF THE LITERATURE ON METACOMPREHENSION AND THE ADULT READER

Cognition is our ability to acquire, store, and use knowledge through a variety of processes such as pattern recognition, attention, short-term and long-term memory, and comprehension (Zabrucky & Cummings, 2004). *Metacognition*, on the other hand, refers to our ability to acquire knowledge of and regulate those cognitive processes (Baker & Brown, 1984; Flavell, 1979). Because reading and learning from text is an essential part of academic achievement (e.g., Lin & Zabrucky, 1998; Maki & McGuire, 2002) and is necessary for overall success in our society (Lorch & van den Broek, 1997), a great deal of attention has been given to understanding metacognition as it relates to text comprehension (*metacomprehension*).

Metacomprehension can be divided into *knowledge* of comprehension and *monitoring* of comprehension (Baker & Brown, 1984). A person's knowledge about comprehension is the relatively stable information and beliefs about what cognitive variables interact to affect the process and outcome of comprehension tasks. It can be loosely divided into three categories consisting of person, task, and strategy variables (Flavell, 1979, 1981). Knowledge about person variables is the information and beliefs a person has acquired with respect to his or her own and others' comprehension abilities (e.g., believing that older adults can improve their comprehension skills). Knowledge about task variables involves what one has learned about the ways in which

comprehension tasks differ from one another and how such differences affect one's success at accomplishing them (Flavell, 1987). An example is knowing that it is easier to remember the gist of a story than to recall it verbatim. Knowledge of strategy variables is a person's repertoire of strategies that can be used to achieve various text processing goals (e.g., rereading to improve understanding).

Knowledge about comprehension is thought to be relatively stable. A person who knows facts and strategies about text processing is likely to continue knowing these facts and strategies across time (Lin & Zabrucky, 1998). Conversely, monitoring of comprehension (or *comprehension monitoring*) is a less stable, on-line activity that an individual may or may not engage in. Monitoring of comprehension consists of two components: evaluation and regulation (Baker, 1985). *Evaluation* of comprehension involves the ability to keep track of how well or how poorly comprehension is proceeding. Readers are evaluating their comprehension when they realize that they do not understand the meaning of a sentence. On the other hand, rereading a sentence in the hopes of understanding its meaning is an example of *regulation*. Regulation of comprehension is the ability to use strategies to resolve comprehension failures that may occur (Baker & Brown, 1984). "Thus, regulation cannot occur without evaluation, but evaluation may occur in the absence of regulation or may not occur at all" (Zabrucky & Cummings, 2004, p. 724).

The primary goal of this paper is to provide an overview of the literature on metacomprehension with respect to the adult reader. Several different research paradigms have been used to study the different aspects of metacomprehension, and this review is organized around those paradigms. Knowledge of comprehension has generally been assessed using a self-report paradigm, primarily with questionnaires and interview

data. Self-reports also have been used to investigate adults' comprehension monitoring by asking readers to report on their behaviors as they read or to give prospective or retrospective accounts of their comprehension processes. The error detection paradigm, sometimes in conjunction with measures of eye movements, also has been used to examine comprehension monitoring. In error detection tasks, various types of errors are inserted into texts to create comprehension failures. Evidence of adults' ability to both evaluate and regulate their comprehension monitoring can be obtained by examining error detection rates and strategy use to repair the comprehension failures. Researchers interested exclusively in adults' evaluation of comprehension have used the calibration paradigm. In calibration studies, participants are asked to read texts and make some type of confidence judgment (e.g., predict their future performance on a comprehension test over the material). By comparing test performance and confidence ratings, researchers have been able to learn how adults evaluate the final state of their comprehension (Lin & Zabrucky, 1998).

In the next sections, these paradigms are explained in detail, and findings from each line of research are presented. Since the majority of participants in the studies reviewed here are college undergraduates, the term "students" is used instead of "participants" and exceptions noted.

Self-Reports

One way to find out about adults' metacomprehension knowledge is to simply ask them. Self-reports have been collected in a variety of ways. Readers have been asked to think-aloud about what they are doing and thinking as they read, to complete questionnaires or checklists about their general strategy use, and to give retrospective reports after completing specific reading tasks.

Self-report data have shown that successful readers use a variety of strategies. Cioffi (1986) asked 43 students identified as good readers by their teachers to read a passage on social anthropology and then complete a reading strategies questionnaire. The questionnaire asked students to report the reading strategies they typically employed as well as the strategies they used to understand the social anthropology passage in particular. The most frequently reported strategies were rereading, looking for main ideas, and stopping to remember what had been read to that point.

In addition to strategies employed by good readers, metacomprehension researchers have been interested in ability-related differences among readers, and self-reports have indicated a number of differences between good and poor readers. Hare (1981) classified students (n = 24) as good or poor readers based on their Nelson-Denny reading scores and asked them to read two articles. The *high-knowledge* article was about teaching practices and written in first person with numerous examples. The *low-knowledge* article was an empirical study written in third person using a technical style. After reading each article, the students were asked to write down everything they noticed about their reading. Good readers reported evaluating their comprehension twice as often as did poor readers when reading both articles. They also reported using a larger number of strategies than poor readers (twice as many when reading the high-knowledge article and three times as many when reading the low-knowledge article). Of additional interest, Hare's good readers reported using strategies not mentioned by poor readers. In particular, good readers reported reading selectively and adjusting reading speed, two strategies that were never mentioned by poor readers.

The finding that good readers use different types of strategies has been supported by other studies. Spring (1985) classified students as good (n = 21) or poor (n = 25)

readers based on their SAT verbal scores and asked them to list the 15 strategies they would typically employ when learning textbook material. A factor analysis of the reading strategies yielded two main factors, which Spring called *comprehension strategies* and *study strategies*. Comprehension strategies included strategies such as relating the material to one's experiences or prior knowledge and mentally identifying the most important ideas in a text. Study strategies, by contrast, were generally rehearsal-type activities such as underlining and taking notes. Spring found that good readers used study strategies as frequently as poor readers but reported using comprehension strategies significantly more often than poor readers. Poor readers, on the other hand, reported that they relied primarily on study strategies to understand a text.

This difference in strategy use was further illustrated in a study conducted by Kaufman, Randlett, and Price (1985). Good (n = 27) and poor (n = 25) readers read short passages from the Davis Reading Test (Form 1A) and completed a questionnaire about their general strategy use. The authors were able to place the reported strategies into three general categories: (1) strategies that affect the learning environment (e.g., cleaning one's room), (2) observable strategies (e.g., rereading), and (3) "inside the head" strategies that are unobservable (e.g., concentrating). Although good and poor readers were equally likely to use Category 1 and 2 strategies, good readers reported using Category 3 strategies significantly more often than poor readers.

From these studies it appears that good and poor readers alike use strategies that aid them in remembering what they have read. Because all of the participants in these studies were college undergraduates, it is reasonable that they would approach reading with the goal of remembering material for an upcoming exam. The good readers in these

studies, however, also used comprehension strategies that would help them understand what they had read.

Given that poor readers tend to adopt strategies that aid them in remembering text rather than comprehending it, is this reflected in their test scores? The answer may depend on the type of question being asked. Walczyk, Marsiglia, Bryan, and Naquin (2001) recorded the think-alouds of 76 students classified as good or poor readers on the basis of four verbal tasks (word-naming, semantic-access, verbal working memory, and anaphor-resolution). The authors were specifically interested in how readers differed in the use of what they referred to as *compensatory behaviors*. The compensatory behaviors consisted of *pauses* (interruptions of reading that did not exceed 5 seconds), *lookbacks* (rereading no more than three words), and *rereading* (rereading four words or more). Walczyk et al. proposed that poorer readers would need to use these behaviors more frequently than good readers because they lacked the verbal ability of better readers. Students were asked to think-aloud as they read a 502-word expository passage and their think-alouds were recorded on audiotape. The authors also asked readers to complete a 16-item comprehension test. As hypothesized, the think aloud data indicated that poorer readers used all three strategies significantly more often than good readers. Despite these efforts, however, the poorer readers did not perform as well on the comprehension test. Although their performance was comparable to that of good readers on factual questions, they performed significantly more poorly on inferential-type questions.

From these findings it is evident that poor readers evaluate their comprehension and attempt to deal with comprehension failures when they encounter them. It is less clear why they do not always succeed. Even though the poor readers in Walczyk et al.'s (2001) study engaged in compensatory behaviors more often than good readers, they

were not able to reach comparable levels of comprehension. As previously discussed, one possible explanation is that poor readers may have different goals for reading and read for the purpose of remembering information rather than comprehending it.

Rather than look at reading ability, Taraban, Rynearson, and Kerr (2000) examined whether differences in academic achievement were related to strategy use. In this study students were asked to complete a strategies questionnaire and to note any additional strategies they might use when encountering comprehension difficulties. Similar to findings from other studies, the questionnaire responses indicated that the most frequently reported strategies were looking for important information, slowing down and rereading, and determining the meaning of unknown words. The authors used cumulative-grade-point average (on a 4-point scale) as a measure of academic achievement and divided students into higher-GPA (M = 3.48) and lower-GPA (M = 2.19) groups. When they examined students' self-reports of strategy use, they found that students in the higher-GPA group reported using significantly more strategies than students in the lower-GPA group and using them significantly more often. Although all students reported using strategies such as rereading or briefly skimming before reading, higher-GPA students reported looking for important information, inferring information, and drawing on prior knowledge more often than lower-GPA students.

The evidence indicates that good and poor readers can be distinguished by the strategies they choose to comprehend texts. Domain expertise also appears to influence strategy use. Lundeberg (1987) compared the verbal reports of experts and novices in the legal field as they analyzed legal cases. The experts were ten law professors and lawyers. The novices were ten adults with minimally a master's degree but no legal experience. Lundeberg was able to identify six general categories of comprehension strategies

reported by the experts: use of context (reading headings, noting dates), overview (summarizing the facts), rereading (selectively rereading terms and facts), underlining (underlining or highlighting important parts of texts), synthesis (merging relevant facts and issues), and evaluation (approving or disapproving of the judge's decision). The only category of strategies used as often by novices as experts was underlining.

In addition to strategy use, self-report data have been collected on other aspects of metacomprehension knowledge. Moore, Zabrucky, Commander, and Morton (1993) developed the Metacomprehension Scale (*MCS*) to assess seven components of metacomprehension. The components are Regulation, Strategy, Task, Capacity, Anxiety, Achievement, and Locus. The Regulation subscale assesses strategy use when comprehension failures are encountered. The Strategy subscale assesses adults' knowledge of techniques to improve comprehension. The Task subscale assesses knowledge of comprehension processes, and the Capacity subscale assesses adults' perception of their own comprehension abilities. The Anxiety subscale assesses feelings of stress related to reading tasks. The Achievement subscale measures the value adults place on good comprehension skills, and the Locus subscale measures the degree to which adults perceive that they can control their comprehension abilities.

Moore, Zabrucky, and Commander (1997a) found that students' scores on the MCS were correlated with their comprehension performance. In this study, 237 students completed the MCS and took a comprehension test that consisted of 15 short expository texts and 60 true-false questions. The authors found that the MCS accounted for 19% of the variance in comprehension performance. High scores on Capacity, Strategy, and Task, together with low scores on Anxiety were associated with improved comprehension performance. In a subsequent study, Moore, Zabrucky, and Commander (1997b) asked

30 younger (*M* = 23.43 years) and 30 older (*M* = 74.50 years) adults to complete the MCS and take a comprehension test after reading 15 short expository passages. Results showed that the MCS accounted for 17.4% of the variance in performance on the comprehension test. In particular, scores on the Regulation and Locus subscales together accounted for 13% of the variance. Adults who reported using strategies when they encountered comprehension failures and believed they had control over their comprehension outperformed others on the comprehension test. No age differences were found with respect to performance on the comprehension test; however, scores on the MCS showed that younger adults reported using strategies significantly more often than older adults and placed a higher value on good comprehension skills.

More recently, however, Lin, Moore, and Zabrucky (2000), found age-related differences in both performance and metacomprehension knowledge. Sixty younger (*M* = 26.63 years) and 60 older (*M* = 70.32 years) adults read two expository passages and two narrative passages, answered 64 multiple-choice questions, and completed the MCS. Results showed that younger adults performed significantly better on the comprehension test. Scores on the MCS accounted for approximately 27% of the variance in younger adults' performance and 26% of the variance in older adults' performance. Of particular interest is that Anxiety was a consistent predictor of performance for both younger and older adults. In fact, it was the best predictor of performance for older adults, whereas scores on the Capacity subscale were the single best predictor of younger adults' comprehension performance. These studies provide strong evidence that metacomprehension knowledge is vital to successful reading comprehension.

To summarize, findings from self-report studies indicate that adult readers are able to evaluate their comprehension during the course of reading and use a variety of reading strategies to resolve comprehension failures when they occur. Poorer readers, however, evaluate their comprehension less frequently than good readers and are less likely to employ effective comprehension strategies when they encounter a comprehension failure. This may be because poorer readers tend to focus on strategies that will help them remember what they have read while good readers employ strategies (such as summarizing important points and relating information in texts to prior knowledge) that promote higher levels of comprehension. Of particular interest is that person variables such as anxiety about reading tasks appear to influence comprehension performance. Findings from the MCS indicate that, in addition to anxiety, adults' beliefs can predict their comprehension performance.

It is important when considering these findings to recognize that adults may not be able to reliably report on the content of their reading processes. Asking readers to think aloud imposes additional burdens on the cognitive processes used for reading and may disrupt normal reading activities (Afflerbach & Johnston, 1984; Garner, 1982). As a result, readers may not report the same processes they would actually use during normal reading. Prospective and retrospective reports, on the other hand, are likely to be less valid than think-alouds because they require readers to retrieve information from long-term memory which can be quite inaccurate (Dunlosky & Hertzog, 2001; Myers, 1991). Readers may not remember what strategies they used or report strategies they did not use (Garner & Alexander, 1989).

Some, but not all, of the evidence suggests that such concerns are warranted. In one study, 50 students read a 1,800-word excerpt from a college text and reported on

their strategy use while they were reading (Brennan, Winograd, Bridge, & Hiebert, 1986). As each student read, an observer recorded the student's behaviors. Of the 10 strategies reported by readers, there was strong agreement between reader reports and observer reports only for underlining. Forty-five students reported that they underlined or highlighted text compared to 42 who were observed doing so. The remaining nine strategies were given high frequency ratings by readers but were rarely observed. For example, 41 students reported that they reread difficult parts of the text, but observers recorded only 16 occurrences of rereading.

It is not always the case, however, that students give inaccurate reports. Garner (1981) reported that students' retrospective reports of strategy use were quite accurate when the reports were made immediately after reading. Twenty students read a short expository text while 20 students recorded their behaviors. Half the readers wrote a summary of the text and recorded everything they remembered about what they did and thought while they were reading. The other half completed the same tasks two days later. The time delay had no effect on readers' ability to summarize the passage but did interfere with the accuracy of their self-reports. The same-day readers reported using an average of 5.3 strategies, and there was almost 100% agreement between readers' reports and observers' reports. Those in the delayed group, however, were able to recall only an average of 1.4 strategies, and the agreement between reader and observer reports dropped to 50%.

Error Detection Paradigm

As previously discussed, successful comprehension monitoring involves both evaluation and regulation of understanding. Evaluation is the ability to recognize comprehension failures during the course of reading, whereas regulation is the ability to

take appropriate actions to resolve comprehension failures when they occur (Baker, 1985; Lin & Zabrucky, 1998). To examine the effectiveness of adults' monitoring, a number of researchers have used the error detection paradigm. The error detection paradigm was developed by Thorndike in 1917 to determine whether children could read texts and then answer questions on what they had read (Brown, 1987). In the error detection paradigm, researchers insert some type of error into texts to disrupt comprehension. The assumption is that if readers are evaluating their comprehension, they will detect the errors because the errors interfere with understanding of the text. Researchers can also use the error detection paradigm to investigate how adults regulate their comprehension by examining which strategies readers use to resolve any comprehension failures they encounter (Lin & Zabrucky).

Baker (1985) reviewed the literature on comprehension monitoring and identified several standards used by adult readers to evaluate their comprehension. The lexical standard involves evaluating the meaning of individual words. The syntactic standard involves evaluating grammatical constraints and would be used to recognized scrambled phrases. The semantic standard involves evaluating text according to one or more subtypes. *Propositional cohesiveness* involves checking that a new proposition is consistent with ones recently encountered in the text. S*tructural cohesiveness* consists of identifying the main theme of a text and relating newly read information to that theme. *External consistency* is checking that the ideas in a text are consistent with one's prior knowledge, and *internal consistency* is checking that the ideas within a text are consistent with one another. *Informational completeness* consists of checking that a text provides all the information necessary to achieve a goal.

To test adults' use of the internal consistency standard, Baker (1979) inserted three types of errors into short expository passages. The errors were inconsistent information (ideas in one sentence conflicted with those in another sentence), unclear reference (the phrase "one type of novel" was inserted in place of "pastoral novel" where three different novel types were under discussion) and inappropriate logical connective (the word "therefore" was substituted for the word "however"). At the first read, students (n = 14) were able to detect only 23% of the errors. When Baker informed the students that the passages contained errors and gave them a description of each kind, the detection rates improved from 23% to 38%. Although all the errors required that students apply the internal consistency standard, some types of errors appeared easier to detect than others. Specifically, students had difficulty identifying inappropriate connectives even after they were given examples.

Although Baker's (1979) students were able to improve their error detection rates when they were given information about the errors, giving students examples does not always improve their performance. Baker (1985) asked students (n = 58) to read passages from college texts that contained nonsense words (to test use of the lexical standard), contradictory information (to test use of the internal consistency standard), or prior knowledge violations (to test use of the external consistency standard). One group was told that there were errors and given examples of each. The other group was told only that the texts contained errors. Results showed that students who had been given examples were no more successful at detecting them than students who knew only that there were errors. Of particular interest was that reading ability appeared to be a factor. Baker (1985) classified the students as good or poor readers based on their SAT verbal scores and found that when poorer readers were given specific examples of the error

types, their detection rates were comparable to that of the good readers who had been told only that the texts contained errors. Again, some errors were more difficult to detect than others. Nonsense words were most likely to be identified, prior knowledge violations less so, and contradictions much less so.

Because some of the students were not warned about the types of errors inserted into texts, the standards they used were indicative of how readers spontaneously evaluate their comprehension. Baker (1985) found that these readers used more than one standard but did not apply all of them. The most commonly used standard was the lexical standard, and this was especially true among the poorer readers. Furthermore, almost half of the poorer readers did not apply the internal consistency standard on any occasion (even when instructed to do so). Baker also found that approximately two thirds of the students never used the external consistency standard, which indicates that they simply accepted the information presented in the text at face value without questioning its accuracy. These findings suggest that readers were primarily concerned with comprehending text at the word level rather than applying standards that require higher levels of processing.

So far the research shows that some comprehension difficulties are easier to detect than others. Nonsense words (requiring use of the syntactic standard) were reported more frequently than errors requiring more extensive text integration processes (internal consistency and external consistency) (Baker, 1985). Readers were better able to detect some errors that required use of the internal consistency standard than others (Baker, 1979). Even when adults were warned about the errors and given examples, however, error detection rates were low.

That adults are poor at error detection tasks is perhaps the most common finding in the literature (Baker, 1979, 1985; Glenberg, Wilkinson, & Epstein, 1982; Yussen & Smith, 1990; Zabrucky, 1990). Many of them also have what Glenberg et al. termed *illusion of knowing*. Illusion of knowing refers to the situation when a reader has failed to detect an error in a passage and simultaneously reports understanding the information in the passage. Glenberg et al. (1982) altered the last sentence of texts to contain either Given or New information. *Given* information was information previously provided in the text and was indicated by using the definite article "the." *New* information, on the other hand, had not been mentioned previously and was indicated by using the indefinite article "a." The authors hypothesized that if students were successfully evaluating their comprehension, they would attempt to match the Given information in the last sentence to information they had previously read and find any contradictions. If the last sentence contained New information, however, students would not reread because there would be no prior information to match, and contradictions would go undetected.

After being forewarned that the texts might contain errors, students (n = 94) read the passages and noted any errors. They then rated their understanding of the text on a scale from 1 (very little understanding) to 4 (understood very well) and answered two true/false comprehension questions. Confirming their hypothesis, Glenberg et al. (1982) found that students detected significantly more errors when the last sentence in the text was marked as Given rather than New. What was more surprising was the number of students who did poorly on the error detection task but reported that they understood the passage. Defining illusion of knowing as failing to detect an error and simultaneously rating one's comprehension as either 3 or 4, the authors found that 31 of the 94 students had illusion of knowing.

Self-report data have provided evidence of ability-related differences in strategy use, and Baker (1985) found that reading ability was related to error detection. Zabrucky (1990) also found such a relationship. Zabrucky replicated Glenberg et al.'s (1982) study to determine if reading ability (assessed by students' scores on the comprehension subscale of the Nelson Denny Reading Test) influenced error detection rates. Consistent with Glenberg et al.'s findings, students (n = 54) in Zabrucky's study detected significantly more contradictions when sentences were marked as Given rather than New. In addition, reading ability was related to both error detection and illusion of knowing. Higher ability readers detected more errors than lower ability readers (62% and 38%, respectively) and exhibited less illusion of knowing.

According to self-reports, rereading is one of the most frequently used strategies when adults encounter comprehension difficulties. In two of the error detection studies discussed (Glenberg et al., 1982; Zabrucky, 1990), students who reread texts in an attempt to match newly introduced information to previously read information were better able to evaluate their comprehension and detected more errors than students who did not reread. Yussen and Smith (1990) reported that rereading improved detection rates but only for one type of error. They altered expository passages to include *general* errors (information in the topic sentence was inconsistent with or contradicted the details in the rest of the paragraph) or *specific* errors (two sentences within a paragraph were inconsistent with or contradicted each other). Half of the 48 students had one opportunity to read the passages, while the other half read the passages two times. Of interest is that rereading improved error detection rates, but only for general errors. Readers who read the passages one time reported an average of 1.69 general errors (range 0-4) and 1.63

specific errors (range 0-4), while readers who reread reported an average of 2.38 general errors and 1.69 specific errors.

In addition to reading ability, motivational orientation may influence monitoring ability. Kroll and Ford (1992) asked students (n = 230) to complete two motivational orientation scales and placed them into highly task oriented or highly ego oriented groups based on their scores. People with a high ego orientation have a desire to appear smart to others and tend to adopt performance goals, while task oriented individuals want to increase their understanding and tend to adopt learning goals. Six weeks later students completed an error detection task in which they read a 6-paragraph passage, rated their understanding, answered two true/false comprehension questions, and reported any errors they had found. Results showed that, regardless of group, error detection rates were low (M = .23, SD = .52, range = 0-3), while self-ratings of comprehension were high (M = 9.66, SD = 1.58, range = 3-12). The authors defined illusion of knowing as the failure to find a contradiction and simultaneously rate comprehension as 3 or 4 (range = 0-4). Kroll and Ford found that 114 of the 230 students exhibited illusion of knowing. They also found that motivational orientation appeared to influence both detection rates and self-assessments of understanding. Students who scored high on ego orientation detected significantly fewer errors than task-oriented students and were significantly more likely to have illusion of knowing.

It is possible that the performance goals associated with ego orientation and the learning goals associated with task orientation are related to what Spring (1985) referred to as study strategies and comprehension strategies. Recall that Spring found that strategies reported by good and poor readers could be placed in two distinct categories, and that good readers used comprehension strategies more frequently while poor readers

relied on study strategies. Additional research will be needed to determine if there is a connection between these two constructs.

In sum, findings from error detection studies indicate that adults are below ceiling level on error detection tasks. Furthermore, a large number of adult readers have what has been termed illusion of knowing--they are poor at detecting errors but self-assess their comprehension as high. Some conditions, such as being forewarned that texts contain errors and being permitted to reread sometimes, but not always, improve error detection rates. Error detection rates also show improvement when the problems require processing at the word level (the lexical standard) rather than more extensive processing (the internal and external consistency standards). In addition, it appears that some individual differences influence error detection rates. Better readers and task-oriented readers were better at reporting errors than poor readers and readers who were ego-oriented.

An important question is whether failing to detect errors in text is analogous to poor monitoring ability. Several researchers contend that there are other explanations for why readers may not report errors (Baker, 1985; Grabe, Antes, Thorson, & Kahn, 1987; Hacker, 1998; Winograd & Johnston, 1982). For example, readers may assign alternative meanings to a text or use prior knowledge to supplement presented information (Winograd & Johnston). The error detection paradigm may be ineffective at determining whether adults are monitoring their comprehension because reading has become such an automatic process for adults that they automatically use fix-up strategies when they encounter errors. Also, failing to detect errors may be the consequence of applying inappropriate evaluation standards rather than a failure to use any standards at all (Baker). For example, a reader who fails to notice a contradiction may not have been evaluating

his or her understanding with respect to an internal consistency standard but may have been using alternative standards.

Others have questioned the ecological validity of the error detection paradigm (Grabe et al.; 1987; Hacker, 1998). They argue that inserting errors into texts alters the normal reading process and that searching for errors is seldom the goal of normal reading. In response to this criticism, it should be noted that researchers have focused on errors that adults (especially students) are, in fact, likely to encounter in everyday life. Whether students are able to detect errors is particularly relevant outside the laboratory because, unfortunately, textbooks do contain errors and coherence problems (Zabrucky, 1990).

Eye Tracking and On-Line Studies

When we read, we make a series of eye movements (*saccades*) separated by periods of time when the eyes are stationary (*fixations*). It is during these fixations that we acquire information from texts (Rayner, 1993). Although most saccades are movements forward, we also make regressive eye movements (sometimes referred to as *lookbacks*) to reread prior material (Raney & Rayner, 1991). Computer-controlled eye tracking procedures allow us to record these eye movements and collect process measures of reading behavior. For example, researchers have found that when readers encounter difficult material, eye tracking data indicate decreases in saccade length together with increases in the duration of fixations and the frequency of lookbacks (Rayner). Also, when readers encounter unfamiliar or unpredictable words, their eyes remain fixated for longer periods of time than when reading familiar or predictable words (Kambe, Rayner, & Duffy, 2001).

Both eye-tracking studies and on-line studies are used to determine readers' behaviors (e.g., frequency of lookbacks, reading time) during the reading process. In eye-tracking studies, participants' eye movements are recorded using specialized equipment. For example, text will be presented on a monitor screen and, as a participant reads the text, the monitor makes a record of the reader's gaze using the relative location of the pupil and corneal reflection. These data then can be used to produce a record of the reader's fixations and saccades (Grabe et al., 1987). In on-line studies, participants' behaviors are recorded as they read text presented on a computer monitor. Readers can initiate presentation of the text on the monitor by pressing different keys (e.g., "next," "back"). The computer records the amount of time each word or sentence is displayed (or redisplayed if a reader presses the "back" button) on the screen.

Eye tracking and on-line studies have been a valuable adjunct to other data collection methods used in metacomprehension research. For example, results from both self-reports and error detection studies indicate that adults often use a rereading strategy when they encounter comprehension problems. Alessi, Anderson, and Goetz (1979) used eye-tracking data to show that rereading does have beneficial effects on comprehension. Students (n = 104) read a 4,926-word expository text displayed on a computer screen and answered multiple-choice questions that were inserted in the text at various intervals. The authors "forced" some students to reread by manipulating the text on the computer screen. When students made an incorrect response to a question, half of them were shown the section of text containing the correct answer on the computer screen. Students who made errors and were permitted to look back at previous sections of the text outperformed those who were not permitted to look back.

Eye tracking data have also been used in conjunction with the error detection paradigm to gather converging information about what occurs when readers encounter comprehension problems. For example, Grabe et al. (1987) asked 40 students to read short expository passages, half of which were altered so that information in one sentence contradicted information in a subsequent sentence. The Uninformed readers were told to read in preparation for a comprehension test. The Informed readers were also warned that some of the paragraphs contained contradictions. The authors reported error detection rates only for the Informed readers and reported that they were able to identify 47% of the errors. Examination of the eye tracking data revealed that at the first read both groups fixated for longer periods of time on the altered sentences than the unaltered sentences. Warning readers about the errors appeared to alter their normal reading processes. The Informed readers engaged in almost four times as many forward and regressive eye fixations as the Uninformed readers when they reread the passages. The authors hypothesized that the difference in reading patterns was due to the Informed readers preparing to report the errors.

In an on-line study, however, Baker and Anderson (1982) did not find that warning readers about errors had a significant effect on their reading behaviors. They used expository passages in which the middle paragraph of each 3-paragraph passage contained an error. The error consisted of replacing one noun or adjective with a word that conveyed an opposite or incompatible meaning. The passages were displayed sentence by sentence on a computer screen, and students (n = 90) could look back at preceding sentences or reread an entire paragraph by pressing computer buttons. Half of the students were warned that the passages might contain errors and were asked to report them. Results showed that, regardless of experimental condition, all readers spent more

time reading inconsistent sentences than consistent sentences and looked back at inconsistent sentences more often. Consistent with results from other studies, the overall detection rate was low (64%) even when readers were warned to look for errors. Furthermore, readers who had been warned to look for errors were no better at detecting errors than Uninformed readers.

Zabrucky and Moore (1994) used on-line measures of reading behavior in conjunction with the error detection paradigm to explore age-related differences in comprehension. Twenty younger (*M* = 22.50 years, range = 19-33) and 20 older (*M* = 71.35 years, range = 62-81) adults read passages from college level textbooks presented one sentence at a time on a computer screen. In each passage, one sentence was changed so that it was either factually consistent or inconsistent with a sentence presented earlier in the text. The authors found no age-related differences with respect to reading speed or error detection rates. Age differences, however, were evident when the authors examined the rereading pattern. The data showed that older adults reread inconsistent sentences as often as consistent sentences, suggesting that they used rereading as a general approach to reading rather than as a strategy to aid in repairing comprehension failures. Younger adults, on the other hand, "selectively" reread the inconsistent sentences, suggesting that they used rereading as a strategy when they encountered errors in the passages. The authors concluded that older adults were as likely as younger adults to evaluate, but not to regulate their understanding.

In a subsequent study of the possible effects of text genre on monitoring, however, Zabrucky and Moore (1999) found that older adults did use rereading as a strategy. Twenty younger (*M* = 22.55 years, range = 18-34) and 20 older (*M* = 69.78 years, range = 61-77) adults read a series of expository and narrative passages presented

one sentence at a time on a computer. Some of the texts contained inconsistent sentences. Zabrucky and Moore found that, regardless of age, adults reread expository passages more frequently than narrative passages and reread inconsistent sentences more frequently than consistent sentences, especially when the inconsistent sentences were in expository texts. They also found that older adults were as likely as younger adults to selectively reread texts. Why did this group of older adults selectively reread? One possible explanation the authors discussed is that the older adults in this study were highly educated and highly verbal compared to those in previous research, which may have mediated any age-related differences in regulation.

In addition to comparisons between younger and older adults, researchers have examined differences between adolescents and adults. In an eye movement study, Grabe, Antes, Kahn, and Kristjanson (1991) compared undergraduates (n = 39) and 6th-7th graders (n = 38). Six paragraphs were adapted from magazines considered appropriate for upper-elementary and junior high school readers. The paragraphs were altered to contain either an internal or an external contradiction. The Uninformed group was told to read in preparation for a comprehension test, and the Informed group was additionally warned that some of the paragraphs contained errors and were given examples of each. The eye movement data showed that, as expected, younger readers displayed slower reading times, and all readers slowed their reading when reading paragraphs that contained contradictions. Informed readers also spent more time rereading the paragraphs than Uninformed readers.

Consistent with the literature, the overall error detection rate was low with participants identifying only 46% of the internal contradictions and 40% of the external contradictions. Of particular interest was the finding that adults significantly

outperformed adolescents only when identifying external inconsistencies. This is particularly surprising because previous research indicated that using the internal consistency standard is relatively difficult for children (Baker, 1985). Grabe et al. (1991) concluded that adolescents were more proficient at error detection than previous research has indicated. However, the reading level of the texts may have been problematic for the adults. As Baker and Brown (1984) have noted, comprehension monitoring is unlikely to occur when texts are too easy because reading progresses automatically. Because the texts in this study were at the high school level, college students may not have monitored their comprehension.

Data from eye-tracking and on-line studies provide some evidence that the presence of errors in texts influences normal reading processes. In all of the studies, data indicated changes in adults' eye movements and reading behaviors when they read problematic texts. However, the low error detection rates suggest that readers were not conscious of the errors. These findings lend support to the view that reading has become such a fluid process for adults that they automatically use fix-up strategies when they encounter errors in texts (Baker, 1985; Baker & Brown, 1984). Nevertheless, it is still not clear why warning adults that texts contained errors and giving them explicit instructions to look for them did not improve error detection accuracy.

Calibration Paradigm

Glenberg and Epstein (1985) used the term *calibration of comprehension* to describe the accuracy with which readers evaluate their comprehension. In the calibration paradigm, readers are asked to read texts and then make a confidence judgment about their future performance on a test over the material (a *prediction*). Alternatively, readers may be asked to assess their prior test performance (a *postdiction*).

Postdictions are sometimes referred to in the literature as *calibration of performance* (Glenberg, Sanocki, Epstein, & Morris, 1987). Calibration accuracy is the relation between the confidence ratings and test performance. The primary finding from a review of the literature is that adults do not evaluate their comprehension as accurately as one might expect. Although results have often indicated calibration accuracy at greater than chance levels, it is generally low with reported gamma correlations in the .20 to .30 range[1]. The major findings of calibration studies discussed in this review are summarized in Table 1.

As can be seen in Table 2, researchers have investigated a number of variables in an effort to learn more about how adults evaluate their comprehension (for a review, see Lin & Zabrucky, 1998). These variables can be classified as subject, task, and text variables. Subject variables include individual differences such as reading ability and domain knowledge that may be related to calibration ability. Task variables are those characteristics of the comprehension test that might influence calibration (e.g., the number of test questions). Text difficulty and genre are examples of text variables that may help account for differences in calibration accuracy. In the following sections, each of these variables will be discussed in turn. Then, ways to improve calibration accuracy will be examined.

---

[1] Goodman and Kruskal's Gamma (*G*) is a nonparametric measure of association employed with ordinal data (Sheskin, 2000). Unlike other correlation coefficients, gamma is not interpreted in terms of variance accounted for, but has a probabilistic interpretation. Calibration accuracy is calculated by computing a gamma correlation between subjective ratings (prediction or postdiction) and comprehension performance. Data from all possible pairs are examined, revealing the probability that a text with a higher comprehension rating also will have a higher performance score. Gamma ranges from –1.00 if higher confidence ratings are paired with lower performance to +1.00 if higher confidence ratings are paired with higher performance.)

Table 1

Summary of Findings from Calibration Studies

Study	*N*	Rating	Major Findings
Commander & Stanwyck, 1997	136	text understanding Likert (1 - 4)	41.9% poor monitors
De Carvalho Filho & Yuzawa, 2001	77	prediction Likert (1 - 5)	23.5% accuracy
Dunlosky & Rawson, 2005	113	prediction Likert (1 - 6)	*G* = .25 and .69*
Glenberg & Epstein, 1985 (Exp. 1)	85	prediction Likert (1 - 6)	r_{pb} = .04 - .07
Glenberg & Epstein, 1985 (Exp. 2)	60	prediction Likert (1 - 6)	r_{pb} = .06 - .12
Glenberg & Epstein, 1985 (Exp. 3)	39	prediction/postdiction Likert (1 - 6)	r_{pb} = .04 pre r_{pb} = 19* - .23* post
Glenberg & Epstein, 1987	57	prediction/postdiction Likert (1 - 6)	*G* = .02 - .06 pre *G* = .36* - .42* post
Glenberg, Sanocki, Epstein, & Morris, 1987 (Exp. 1)	63	prediction Likert (1 - 6)	*G* = .11
Glenberg et al., 1987 (Exp. 2)	34	prediction Likert (1 - 6)	*G* = .24
Glenberg et al., 1987 (Exp. 6)	46	prediction Likert (1 - 6)	*r* = .13*
Glenberg et al., 1987 (Exp. 7)	38	prediction Likert (1 - 6)	*G* = .40*
Glenberg et al., 1987 (Exp. 8)	37	prediction Likert (1 - 6)	*G* = .35*

(*table continues*)

Study	*N*	Rating	Major Findings
Keleman, Frost, & Weaver, 2000 (Exp. 1)	66	prediction Likert (1 - 6)	*G* = .38* and .44*
Keleman et al, 2000 (Exp. 2)	78	prediction Likert (1 - 6)	*G* = .02 and .05
Lin, Moore, & Zabrucky, 2000	120	postdiction Likert (1 - 6)	significant *r* with knowledge scale
Lin, Moore, & Zabrucky, 2001	60	prediction/postdiction Likert (1 - 7)	*G* = .14* pre *G* = .28* post
Lin, Zabrucky, & Moore, 1997	31	prediction Likert (1 - 6)	*r* = .09
Lin, Zabrucky, & Moore, 2002	111	text comprehension Likert (1 - 6)	*G* = .09 - .47*
Lundeberg, Fox, & Puncochar, 1994	254	postdiction Likert (1 - 5)	83.0 % accuracy
Magliano, Little, & Graesser, 1993 (Exp. 1)	63	text comprehension Likert (1 - 6)	*r* = .11 - .54*
Magliano et al., 1993 (Exp. 2)	145	text comprehension Likert (1 - 6)	*r* = .16 - .53*
Maki, 1995	54	prediction/postdiction Likert (1 - 6)	*G* = .01 - .27* pre *G* = .22 - .69* post
Maki & Berry, 1984 (Exp. 1)	30	prediction Likert (1 - 6)	*r* = -.03 - .15*
Maki & Berry, 1984 (Exp. 2)	39	prediction Likert (1 - 6)	*r* = -.05 - .23*
Maki, Foley, Kajer, Thompson, & Willert, 1990 (Exp. 1)	75	comprehension ease or prediction, postdiction Likert (1 - 6)	*G* = .13* - .32* pre *G* = .45* - .59* post

(*table continues*)

Study	*N*	Rating	Major Findings
Maki et al., 1990 (Exp. 2)	39	comprehension ease or prediction, postdiction Likert (1 - 6)	*G* = .18* - .37* pre *G* = .52* - .53* post
Maki, Jonas, & Kallod, 1994	71	prediction/postdiction Likert (1 - 6)	*G* = .11 pre *G* = .55* post
Maki & Serra, 1992a (Exp. 1)	44	text comprehension or prediction, postdiction Likert (1 - 6)	*G* = .17* - .31* pre *G* = .28* - .40* post
Maki & Serra, 1992a (Exp. 2)	45	text comprehension or prediction, postdiction Likert (1 - 6)	*G* = .14 - .40* pre *G* = .34* - .41* post
Maki & Serra, 1992a (Exp. 3)	69	text comprehension or prediction, postdiction Likert (1 - 6)	*G* = -.04 - .25* pre *G* = .09 - .31* post
Maki & Serra, 1992b (Exp. 1)	71	prediction/postdiction Likert (1 - 7)	*G* = .22* - .23* pre *G* = .32* - .35* post
Maki & Serra, 1992b (Exp. 2)	71	prediction/postdiction Likert (1 - 7)	*G* = .11 - .32* pre *G* = ..24* - .45* post
Maki & Serra, 1992b (Exp. 3)	36	prediction/postdiction Likert (1 - 7)	*G* = .20 - .38* pre *G* = .21* - .37* post
Moore, Lin, & Zabrucky, 2005	60	prediction/postdiction Likert (1 - 7)	r = .66* pre r = .65* post
Palmer, 1995	92	text comprehension Likert (1 - 4)	*r* = .03 and .26 *r* = .08 and .82*
Pierce & Smith, 2001 (Exp. 1)	44	prediction/postdiction (total correct items)	*G* = -.44* - .14 pre *G* = .34* - .58* post

(*table continues*)

Study	*N*	Rating	Major Findings
Pierce & Smith, 2001 (Exp. 2)	49	prediction/postdiction (total correct items)	*G* = -.20 - .23 pre *G* = .31* - .41* post
Pressley & Ghatala, 1988	45	postdiction Likert (1 - 6)	*G* = .38*
Pressley, Ghatala, Woloshyn, & Pirie, 1990	48	postdiction Likert (1 - 7)	*ns* confidence for correct/incorrect items
Rawson, Dunlosky, & McDonald, 2002 (Exp. 1)	160	text comprehension/ prediction Likert (1 – 6)	*G* = .18 - .40
Rawson, et al., 2002 (Exp. 2)	80	text comprehension/ prediction Likert (1 - 6)	*G* = .12 - .20
Rawson, Dunlosky, & Thiede, 2000 (Exp. 1)	75	prediction (0% - 100%)	*G* = .24 - .57*
Rawson et al., 2000 (Exp. 2)	56	text understanding Likert (1 - 7)	*G* = .11 - .55*
Schommer, 1990	266	text understanding Likert (1 - 4)	significant *r* with knowledge scale
Schommer, Crouse, & Rhodes, 1992	424	text understanding Likert (1 - 4)	significant *r* with knowledge scale
Schommer & Surber, 1986	48	prediction Likert (1 - 4)	*r* = .23 - .27
Schraw, 1994	115	postdiction Likert (1 - 5)	23.9% accuracy
Schraw, 1997	95	postdiction (0% - 100%)	*r* = .30*

(*table continues*)

Study	*N*	Rating	Major Findings
Schraw, Dunkle, Bendixen, & Roedel, 1995	134	postdiction (0% - 100%)	multiple measures
Schraw, Potenza, & Nebelsick-Gullet, 1993	85	postdiction (0% - 100%)	r_{pb} = .05 - .33
Walcyzk & Hall, 1989	60	prediction/postdiction Likert (1 - 6)	*r* = .30 - .58* pre *r* = .02 - .63* post
Weaver, 1990 (Exp. 1)	21	prediction Likert (1 - 6)	*r* = .10
Weaver, 1990 (Exp. 2)	54	prediction Likert (1 - 6)	*r* = .08 - .32*
Weaver & Bryant, 1995, (Exp. 1)	92	prediction Likert (1 - 6)	*G* = .21* - .32*
Weaver & Bryant, 1995, (Exp. 2)	91	prediction Likert (1 - 6)	*G* = .29* - .69*

Note. All findings are group means.

G = gamma correlation, *r* = Pearson's correlation; r_{pb} = point biserial correlation,

pre = prediction, post = postdiction

*$p \leq .05$

Table 2

Features of Calibration Studies

Study	Variables	Materials
Commander & Stanwyck 1997	reading ability, passage length	1 expository text 3 m-c items
De Carvalho Filho & Yuzawa, 2001	effect of social cues, metacognitive knowledge	6 narrative texts 4 m-c items per text
Dunlosky & Rawson, 2005	rereading	6 expository texts 6 m-c items per text
Glenberg & Epstein, 1985 (Exp. 1)	time between reading and prediction	15 expository texts 1 t-f item per text
Glenberg & Epstein, 1985 (Exp. 2)	practice	15 expository texts 1 t-f item per text
Glenberg & Epstein, 1985 (Exp. 3)	practice, number of test items	15 expository texts 1 t-f item per text
Glenberg & Epstein, 1987	prior knowledge	16 expository texts 1 t-f item per text
Glenberg, Sanocki, Epstein, & Morris, 1987 (Exp. 1)	questions for inference or verbatim recognition, test delay	15 expository texts 1 t-f item per text
Glenberg et al., 1987 (Exp. 2)	test delay	15 expository texts 4 f-c items
Glenberg et al., 1987 (Exp. 6-8)	type of feedback	16 expository texts 3 f-c items
Keleman, Frost, & Weaver, 2000 (Exp. 1)	general monitoring ability	4 narrative texts 10 m-c items per text

(*table continues*)

Study	Variables	Materials
Keleman et al., 2000 (Exp. 2)	general monitoring ability	15 narrative texts 1 m-c item per text
Lin, Moore, & Zabrucky, 2000	age, text genre, metacognitive knowledge	2 narrative and 2 expository texts 16 m-c items per text
Lin, Moore, & Zabrucky, 2001	calibration across multiple measures	12 expository texts 4 t-f items per text
Lin, Zabrucky, & Moore, 1997	text interest	15 expository texts 4 t-f items per text
Lin, Zabrucky, & Moore, 2002	text difficulty	2 narrative & 2 expository texts 12 m-c items per text
Lundeberg, Fox, & Puncochar, 1994	gender, prediction	2 exams 50 m-c items
Magliano, Little, & Graesser, 1993	shallow/deep processing, test item format	8 expository texts 16 t-f or 20 s-a per text
Maki, 1995	questions for important/ unimportant information	12 expository texts 9 m-c items per text
Maki & Berry, 1984 (Exp. 1)	feedback	2 expository texts 18 m-c items per text
Maki & Berry, 1984 (Exp. 2)	test delay	2 expository texts 18 m-c items per text
Maki, Foley, Kajer, Thompson, & Willert, 1990 (Exp. 1)	shallow/deep processing	1 expository text 48 s-a items
Maki et al., 1990 (Exp. 2)	shallow/deep processing	1 expository text 30 s-a items

(*table continues*)

Study	Variables	Materials
Maki, Jonas, & Kallod, 1994	reading ability	1 expository text 48 m-c items
Maki & Serra, 1992a (Exp. 1)	basis for text predictions	12 expository texts 4 m-c items per text
Maki & Serra, 1992a (Exp. 2)	prior knowledge, text difficulty	12 expository texts 4 m-c items per text
Maki & Serra, 1992a (Exp. 3)	prior knowledge, text difficulty	12 expository texts 6 m-c items per text
Maki & Serra, 1992b (Exp. 1)	practice, text difficulty	12 expository texts 4 m-c items per text
Maki & Serra, 1992b (Exp. 2)	practice, text difficulty	12 expository texts 3 m-c items per text
Maki & Serra, 1992b (Exp. 3)	feedback	12 expository texts 3 m-c items per text
Moore, Lin, & Zabrucky, 2005	basis for calibration judgments	12 expository texts 4 t-f items per text
Palmer, 1995	calibration of college freshmen and seniors	1 narrative text and 1 expository text 12 open-ended items
Pierce & Smith, 2001 (Exp. 1)	prediction/postdiction, practice	4 narrative texts 16 m-c items per text
Pierce & Smith, 2001 (Exp. 2)	prediction/postdiction, practice	3 expository texts 16 m-c items per text
Pressley & Ghatala, 1988	item difficulty, reading ability	3 expository texts 5 m-c items per text
Pressley, Ghatala, Woloshyn, & Pirie, 1990	questions for main points or details of text, text difficulty	20 texts 1 m-c item per text

(*table continues*)

Study	Variables	Materials
Rawson, Dunlosky, & McDonald, 2002 (Exp. 1)	test delay	13 expository texts 2 t-f items per text
Rawson, et al., 2002 (Exp. 2)	test delay	13 expository texts 2 t-f items per text
Rawson, Dunlosky, & Thiede, 2000 (Exp. 1)	rereading	6 expository texts 6 m-c items per text
Rawson et al., 2000 (Exp. 2)	rereading	7 expository texts 6 m-c items per text
Schommer (1990)	epistemological beliefs	1 expository text 10 m-c items per text
Schommer, Crouse, & Rhodes, 1992	epistemological beliefs	2 expository texts 15 m-c items per text
Schommer & Surber, 1986	text difficulty	2 expository texts 3 m-c items per text
Schraw, 1994	metacognitive knowledge	7 expository texts 4 m-c items per text
Schraw, 1997	metacognitive knowledge	3 expository texts 4 m-c items per text
Schraw, Dunkle, Bendixen, & Roedel, 1995	general monitoring ability	8 tests in different domains 6 to 28 m-c items per test
Schraw, Potenza, & Nebelsick-Gullet, 1993	feedback, incentives, text difficulty	8 expository texts 36 m-c items math probability test 8 m-c items
Walcyzk & Hall, 1989	examples and questions in text	1 expository text 25 m-c items

(*table continues*)

Study	Variables	Materials
Weaver, 1990 (Exp. 1)	prior knowledge	14 expository texts 1 t-f item per text
Weaver, 1990 (Exp. 2)	number of test items	14 expository texts 1, 2, or 4 t-f item(s) per text
Weaver & Bryant, 1995 (Exp. 1)	text genre	4 narrative and 4 expository texts 16 m-c items per text
Weaver & Bryant, 1995 (Exp. 2)	text difficulty	12 expository texts 16 m-c items per text

Note. f-c = forced choice, m-c = multiple-choice, t-f = true/false; s-a = short answer

Subject Variables

As Lin and Zabrucky (1998) note, there are reader characteristics that may influence both comprehension and metacomprehension. These variables cover a range of individual differences such as knowledge about reading processes, reading ability, and epistemological beliefs. In addition, other **"**person**"** variables such as affective and social influences may impact calibration performance.

Reading ability. A number of differences have been found between good and poor readers with respect to strategy use and error detection rates. Maki and Berry (1984) found that differences in reading ability also influenced calibration accuracy. They asked 30 students to read a half-chapter of an introductory psychology text, after which they rated how well they thought they would do on a subsequent comprehension test. Students returned the next day and answered multiple choice questions on the half-

chapter they had read the previous day. This procedure (reading a half-chapter, making one or more predictions, and returning the next day for the comprehension test) was then repeated. Results showed that students who scored above the median on the first comprehension test were more confident when they were right and less confident when they were wrong, indicating they had less illusion of knowing than those students who had scored below the median. In a second experiment, Maki and Berry asked students to make predictions immediately after reading. This time students who scored below the median were as accurate at calibrating their performance as students who scored above the median.

Maki and Berry (1984) used a median split on a comprehension test to determine reading ability. Others have used a variety of measures to calculate reading ability and have failed to find a relationship between reading ability and calibration accuracy. Pressley and Ghatala (1988) used SAT verbal scores as an independent assessment of reading ability. In this study, students read several short expository texts. After reading each text, they answered test questions, were shown the questions again, and asked to rate their confidence in having made the correct response (20% = "just a guess" to 100% = "absolutely certain"). The authors found no correlation between reading ability and postdiction accuracy. Postdictions were low but greater than chance (mean G = .38).

Commander and Stanwyck (1997) were also unable to find a relationship between reading ability and calibration accuracy. They classified students as good and poor readers based on their Nelson Denny Reading Test (Form C) scores. Students read expository passages from introductory college texts, made predictions, answered comprehension questions, and then made postdictions. The authors combined the

prediction and postdiction ratings and found that reading ability had no effect on calibration accuracy.

Maki, Jonas, and Kallod (1994) determined students' reading ability by scores on the Multi-Media Comprehension Battery (*MMCB*), which assesses comprehension of written, auditory, and pictorial stories, and the Nelson-Denny reading test (Form C) which assesses comprehension and reading rate. Again, reading ability was not related to calibration performance. Although students were able to predict their performance at a greater than chance level, accuracy was quite low (mean $G = .11$). The only score that correlated significantly with prediction accuracy was the auditory comprehension portion of the MMCB.

Prior knowledge. In an early study, Glenberg and Epstein (1987) tested the hypothesis that readers use their prior knowledge about a domain when making predictions about their comprehension performance (the *domain familiarity hypothesis*). Physics and music students read both physics and music texts, made predictions, answered comprehension questions, and then made postdictions. The authors computed separate gammas for the physics and music texts and found that physics students' predictions were negatively correlated with the number of physics courses they had taken. Students' postdictions, however, were at better than chance levels and unrelated to the number of physics or music courses taken. The authors proposed that students based their predictions on their background knowledge of a topic rather than their comprehension of the texts but based their postdictions on how well they had comprehended the texts.

Further support for the domain familiarity hypothesis came from Glenberg et al. (1987). Students read 15 one-paragraph texts after which they rated their familiarity with

the material in the texts, made a prediction rating, and answered a true-false question. Students' domain familiarity ratings were correlated with their prediction ratings but not with performance on the comprehension test. Other findings, however, suggest that adults do, in fact, use information from texts when predicting their performance on an upcoming test. Maki and Serra (1992a, Exp. 2) showed students the titles of texts with short descriptor sentences and asked them to rank order how difficult they thought the texts would be to comprehend or how well they thought they would do on a test of the material. The authors used the ranks as a measure of students' prior knowledge. Students then read four texts, after which they rank ordered the texts again and took a short comprehension test. Maki and Serra found no relationship between the prereading ranks and subsequent test performance. After reading the texts, however, students were able to predict their performance at greater-than-chance levels, suggesting that they gained knowledge from reading the texts and were able to use this knowledge to accurately predict their test performance.

The finding that readers base their predictions on material in the texts was further supported by Weaver (1990). In this study, students read 12 short expository texts, made a prediction, and then answered 12 true-false questions. A control group took the comprehension test without reading the texts. The experimental group answered 79% of the questions correctly compared to 57% for the control group. Weaver eliminated five questions that had been answered with greater than 75% accuracy by the control group because the questions could be answered on the basis of world knowledge. Prediction accuracy for the remaining questions was no better than chance (r = .15). Because the remaining questions could not be answered on the basis of prior knowledge, Weaver

concluded that students had based their predictions on their comprehension of the texts rather than on preexisting world knowledge.

Interest. Research has shown that when readers are interested in a topic, they tend to comprehend a text at a deeper level (Schiefele, 1992). However, we know little about the possible influence of interest on calibration. Lin, Zabrucky, and Moore (1997) investigated the effects of both *domain* interest (students' interest in a particular subject) and *specific* interest (how interesting students found a particular passage) on calibration accuracy. Students were asked to report their interest in eight subject areas (political science, physics, philosophy, biology, economics, engineering, geography, and astronomy). They then read texts in each of the areas, rated how interesting they found them, and made a prediction about their performance on a comprehension test. Students who reported higher levels of specific interest scored higher on the comprehension test and showed higher levels of calibration accuracy. Domain interest, however, was not related to either test performance or calibration accuracy.

Gender. Although no studies have been found that investigated gender differences on calibration in lab settings, Pallier (2003) found that gender influenced confidence ratings on several cognitive tasks. In the first study, 185 students took a general knowledge test (20 questions that assessed knowledge of history, geography, current events, science, and technology) and a visualization test (a line length test in which five vertical, nonaligned lines were presented simultaneously on a computer screen, and students had to identify the longest line). Students were asked to make a confidence judgment after each test item. Although there was virtually no difference between the accuracy scores of men and women, men exhibited higher confidence on both tasks compared to women.

In a second study, 303 men and women ranging in age from 17 to 80 years (M = 29.11; SD = 12.85) were asked to complete tests of crystallized and fluid intelligence (Pallier, 2003). Again, test performance showed no significant gender differences. However, men's confidence ratings were significantly higher than women's confidence ratings. As Pallier pointed out, not only do men report consistently higher confidence in their cognitive ability, but this confidence appears to remain stable across the lifespan.

In another related study, students made postdictions after taking exams in a laboratory methods course and a memory course (Lundeberg, Fox, & Puncochar, 1994). Results showed no gender differences on postdiction accuracy for the memory course exam, but a gender difference was evident on the postdictions for the lab exams. The women's postdictions were significantly more accurate than the men's (87% and 37% respectively). Also, the males exhibited more illusion of knowing. When men made incorrect responses, their confidence ratings were close to 4 (on a 5-point scale), but when women were incorrect, their confidence ratings were closer to 3.

Social Influences. We know very little about how social influences affect calibration accuracy. In an error detection study, Kroll and Ford (1992) found that students who wanted to increase their understanding detected more errors while students who had a desire to outperform others and to appear smart were more likely to have illusion of knowing.

Related research also suggests that there might be a relationship between social influences and calibration accuracy. Karabenick (1996) found that students were more likely to reveal they were confused about a videotaped message when they thought others had also expressed confusion. Karabenick informed students that they were going to be assessed on their ability to understand the contents of two 8-minute videotaped messages

that contained information from news articles about environmental issues. To ensure that the messages would be difficult to understand, the presenter mispronounced words, omitted sentences, spoke too softly to be clearly heard, and rearranged words. Students watched one videotape alone and one videotape when they thought other students (simulated) had questions about the videotape. Results of three experiments revealed that when students thought others had questions about the videotapes, they were more likely to report their own confusion.

De Carvalho Filho and Yuzawa (2001) found that, in some cases, students' calibration judgments could also be influenced by the behavior of others. Students at Hiroshima University completed a checklist of their strategy knowledge (the General Monitoring Strategies Checklist; *GMSC*, Schraw, 1997) and were assigned to high- or low-knowledge groups based on their scores. Students then read four narrative texts, took a comprehension test, and made postdictions. The authors divided the students into high- and low-monitor groups based on their postdiction accuracy. On the sheets provided for the postdictions, the authors added social cues about how students in a previous fictitious study had answered the same questions. One group read that students had done well, while a second group read that students had done poorly. A control group received no social cues. Results showed no relationship between social cues and metacognitive knowledge. Although high monitors were not affected by the social cues, low monitors gave higher postdictions when the cues indicated that others had done well, and decreased them when they thought others had done poorly.

Metacognitive knowledge. One's epistemological beliefs or views about the nature of knowledge and learning are a part of stable metacognitive knowledge. There is some evidence to suggest that these beliefs influence calibration accuracy. Schommer

(1990) asked students to complete a questionnaire about their beliefs and then read a passage from a psychology text, after which they rated their understanding of the passage. The questionnaire was composed of 12 subsets of items that assessed views of learning and knowledge. Students completed the tasks at home and took a comprehension test at the next class meeting. A factor analysis of the questionnaire items resulted in four main factors: Innate Ability (e.g., "Ability to learn is innate."), Simple Knowledge (e.g., "Knowledge is discrete and unambiguous."), Quick Learning (e.g., "Successful students learn things quickly."), and Certain Knowledge (e.g., "Scientists can ultimately get to the truth.") (p. 500). Schommer then compared students' test performance and confidence ratings with their scores on the epistemological questionnaire. Results showed that the Quick Learning factor predicted both test performance and overconfidence. Students who believed that they should learn without too much time or effort did not perform as well as others on the comprehension test and were more likely to overestimate their understanding of the passage.

In a follow-up study, Schommer, Crouse, and Rhodes (1992) asked undergraduate and graduate students to complete Schommer's (1990) epistemological questionnaire, after which they read passages from an introductory statistics text, made predictions, and took a comprehension test. A factor analysis resulted in four main factors, with the factor structure almost identical to that obtained by Schommer. Similar to Schommer's results, one factor predicted both test performance and overconfidence. This time, however, the factor was Simple Knowledge. Students with high scores on Simple Knowledge (e.g., reported that they believed knowledge is merely isolated facts, or that words have only one meaning) had the poorest performances on the comprehension test and were more overconfident about their understanding of the texts. It is not clear why Quick Learning

predicted test performance and overconfidence in one study, while Simple Knowledge predicted test performance and overconfidence in the second study. There are several possible explanations. In Schommer's study, students rated their understanding of the passage, but made predictions about their future test performance in Schommer et al.'s study. This difference in ratings may have influenced students' confidence levels so that students with high scores on the Quick Learning dimension felt more overconfident when rating their understanding than when making predictions, whereas students with high scores on Simple Knowledge showed the opposite pattern. A second possibility is that the beliefs measured by the Simple Knowledge and Quick Learning factors are similar in some respects. While the results of these two studies clearly support the hypothesis that epistemological beliefs are related to calibration accuracy, further research is needed to get a clearer picture of the relationship.

In addition to beliefs about knowledge and learning, metacognitive knowledge involves people's perceptions of themselves as learners. Schraw (1994) found that students' self-perceptions were related to their postdiction accuracy. He asked students to assess their own monitoring ability by marking a slash on a 100-millimeter line that ranged from 0 (poor monitoring ability) to 100 (excellent monitoring ability) and classified them as high-, average- or low-monitors based on their self-reports. Students then read seven expository texts taken from the Nelson-Denny Reading Test (Form E). After reading each text, students answered multiple-choice questions and made postdictions. Similar to findings from other studies, calibration accuracy was significant but low (r = .29). Schraw found that students who assessed themselves as high monitors outperformed the average and low monitors on the comprehension test and were more

accurate in their postdictions. These findings provide further evidence that metacognitive knowledge is related to both performance and calibration of performance.

General metacognitive ability. Most calibration studies have examined adults' ability to effectively monitor their performance specifically in the domain of reading comprehension. Schraw, Dunkle, Bendixen, and Roedel (1995) compared students' performance in a variety of domains to determine if monitoring ability is tied to expertise in a domain or is a more general metacognitive ability. According to the *domain-specific hypothesis*, monitoring ability is directly related to one's knowledge in a domain area. If, for example, a student's skills in mathematics improved, he or she would show improvement at monitoring performance on a mathematics test, but would not show improved monitoring in other content areas. Conversely, the *domain-general hypothesis* proposes that metacognitive knowledge is qualitatively different from other types of domain knowledge and can be used to monitor performance across content domains. When monitoring their performance on a mathematics test, for example, students will rely on content knowledge but will also rely on general metacognitive knowledge (e.g., their knowledge of themselves as learners, of cognition, of test-taking strategies such as checking comprehension of test questions). As they acquire more metacognitive knowledge, their ability to monitor their performance in any content domain will improve as well.

If adults possess a general metacognitive ability as proposed by the domain-general hypothesis, they should show consistent levels of monitoring accuracy across tasks even when performance scores are not consistent (Maki & McGuire, 2002). To test this, Schraw et al. (1995) compared students' monitoring ability in eight domains (reading comprehension, geographical distances between American cities, calorie value

of common foods, U. S. Presidents, running speed of animals, spatial judgments, general knowledge, and mathematical word problems). Rather than assess calibration accuracy using correlational methods, the authors computed measures of bias and discrimination. Discrimination is calculated by subtracting the average confidence rating for incorrect items from the average confidence rating for correct items. The scores are then converted to proportions so that the resulting index ranges from +1 to -1. A discrimination score greater than zero indicates that a student correctly assigned high confidence to correct items and low confidence to incorrect items, whereas a score less than zero indicates confidence is low for correct responses but high for incorrect responses (Maki & McGuire). The bias index measures a student's overall overconfidence or underconfidence. Bias scores are derived by obtaining the signed difference between students' average confidence and average performance for each test. Bias scores range from +1 to -1, with scores greater than zero reflecting overconfidence, and scores less than zero reflecting underconfidence (Keleman, Frost, & Weaver, 2000).

Schraw et al. (1995) found that performance scores were not correlated across the domains indicating that students were not equally proficient in all domain areas. If a general monitoring skill exists, the majority of discrimination scores should be correlated even though performance varied across domains. This would indicate that the tendency to make accurate confidence judgments in one domain is related to the tendency to make accurate confidence judgments in other domains. Instead, results supported a domain-specific hypothesis in that discrimination scores were not correlated but varied as a result of performance. Students who had higher performance scores in a domain were also better able to monitor their performance in that domain.

A comparison of the bias scores offered some support for the domain-general hypothesis. The data from the bias scores revealed that students tended to be uniformly over- or underconfident about their performance even though their performance varied across the eight tests. Specifically, all but 2 of the 28 correlations reached significance. These findings suggest that bias (overconfidence or underconfidence) may represent a general metacognitive characteristic, but the ability to accurately monitor one's performance is more likely to be domain specific.

In a later study, Schraw (1997) developed a 10-item rating scale (General Monitoring Strategies Checklist, *GMSC*). Examples of items on the GMSC are "I ask myself periodically if I am doing well" and "I am aware of what strategies I use when I solve problems" (p. 146). Students took tests in four domains (reading comprehension, lexical judgments, mathematical reasoning, and syllogistic reasoning) and made confidence judgments about their performance after each test. Schraw found that performance was not consistent across the four domains, but that all four sets of confidence judgments correlated significantly with GMSC scores. This finding suggests that students used their general metacognitive knowledge when monitoring their test performance (the domain-general hypothesis).

Reporting contradictory results, Keleman et al. (2000) found that general metacognitive ability was not related to increased monitoring accuracy when they compared students' performance on four metacognitive tasks: ease of learning judgments (*EOL*s), judgments of learning (*JOL*s), feelings of knowing (*FOK*s) (Nelson & Narens, 1990, 1994), and comprehension monitoring. The EOL task consisted of recalling 15 pairs of Swahili-English word pairs. The JOL task consisted of recalling 36 unrelated pairs of English concrete nouns. The FOK task consisted of answering 25 general

knowledge questions. For the reading comprehension task, students read four short narrative texts, made a prediction, and then answered 40 multiple-choice questions. The authors found that performance was correlated across the tasks over 60% of the time, but monitoring accuracy was correlated across tasks only 10% of the time. Although students' performance was not consistent, their confidence ratings were consistent in magnitude. In other words, they tended to be reliably underconfident or overconfident about their performance on the four tasks.

Stable trait of confidence bias. Schraw et al.'s (1995) and Keleman et al.'s (2000) findings suggest that adults may possess a general confidence factor. Some evidence for this comes from researchers in differential psychology who have examined confidence bias. *Confidence bias* refers to systematic errors that individuals make when reporting their confidence on intellectual or perceptual tasks (Pallier, 2003). Pallier et al. (2002) investigated confidence bias with a sample of 520 United States Air Force recruits. The recruits took eight tests in four cognitive domains (general knowledge, vocabulary, visualization, and abstract reasoning) and made confidence judgments about their performance. The results of a confirmatory factor analysis showed that confidence ratings across all eight tests loaded on a single factor, indicating that the recruits were systematically under- or overconfident about their performance. Results further showed that overconfidence was generally associated with lower test scores. Individuals who performed poorly on the tests were more likely to be overconfident than those who performed well.

In a recent calibration study, Moore, Lin, and Zabrucky (2005) found further evidence of confidence bias. Students read 12 short expository texts that were divided into three difficulty levels (10th grade reading level, 13th grade reading level, and 16th

grade reading level). The authors hypothesized that students would base their prediction ratings on their general reading ability rather than on their test performance. After reading each text, students were asked to rate their confidence about answering questions on a subsequent comprehension test, how easy they found the text, and how well they understood the text. They then answered four inference questions and made a postdiction. When collapsed across all 12 texts, performance scores and calibration judgments were highly correlated. Thus, students were fairly good at assessing their overall performance. When looking at the postdictions for each text; however, the authors found that the postdictions were not independent across trials. Had students based their calibration judgments on their comprehension of individual passages, the judgments should have varied as a result of the differences in text difficulty. Instead, Moore et al. found that students' evaluation of their performance could be predicted from their evaluation of previous texts. This suggests that students based their calibration judgments on their self-assessments of reading ability rather than on the specific comprehension tasks.

To summarize, it does not appear that adults possess a general metacognitive ability that leads to improved monitoring accuracy. However, there is evidence that some adults base their monitoring judgments on inaccurate beliefs about themselves as learners. When adults rely on this knowledge, it can lead to a consistent pattern of overconfidence or underconfidence when monitoring performance on cognitive tasks.

Task Variables

In addition to reader characteristics, a complete picture of calibration ability requires consideration of task variables such as the difficulty level of the test, test format, and retention interval demands. Issues such as reliability of the comprehension test and

measures used to assess calibration are additional factors that may impact calibration accuracy. Also, asking adults to make predictions may require different, and perhaps more difficult, metacognitive judgments than asking them to make postdictions.

Type of information. Pressley, Ghatala, Woloshyn, and Pirie (1990) examined whether calibration accuracy varied as a result of the type of information requested on the comprehension test. Students read 21 expository passages from the PSAT and SAT and answered multiple choice and short answer questions regarding the main theme of the passage or specific details. Instead of asking students for a written postdiction, the authors instructed students to reread if they thought they had made an incorrect response. If students reread a passage after answering an item, the authors concluded that they were not confident about their response. The data showed that students were significantly more likely to reread a passage after they had incorrectly answered short answer items compared to multiple choice items and were more likely to reread if the question pertained to specific details rather than the overall theme. After comparing the results of the comprehension test and the rereading behaviors, the authors concluded that students were better at assessing their performance when questions related to details of a text rather than the overall theme.

Glenberg et al. (1987), however, found that information type had no effect on students' prediction accuracy. In the first experiment, students read 15 one-paragraph expository texts, made predictions, and answered questions that asked for either inference or factual recall. Results showed that students' prediction accuracy was no greater than chance regardless of information type. In a second experiment, students were asked to respond to either inferential questions or questions that were close paraphrases of ideas in the texts. Again, information type did not influence students' prediction accuracy.

Maki (1995) obtained mixed results when comparing students' prediction and postdiction accuracy on questions that asked for three types of information (important, unimportant, and higher order). Important and unimportant information came directly from the text whereas higher order information required that students make inferences about information in the texts. Expository texts were presented one sentence at a time on a computer screen, then repeated. Although students' test performance was significantly better for the important questions (80% correct) and the unimportant questions (75% correct) than for the higher order questions (64% correct), information type did not influence either prediction or postdiction accuracy. In a second experiment, however, students' prediction accuracy was greater when questions required students to recall unimportant or higher order information.

From these findings, it is not possible to get a clear picture of the effects of information type on calibration accuracy. Further research needs to be conducted to determine the reason for the different outcomes.

Test difficulty. Information type appears to have a mixed effect on calibration accuracy. This may be because items that require students to retrieve factual information may be easier (or more difficult) than inferential questions. Only one study has been conducted that directly investigated the effect of test difficulty on calibration. Pressley and Ghatala (1988) asked students to read several short expository texts and answer multiple-choice questions that asked them to identify the author's purpose, the main idea of the passage, or what inferences might follow from the passage. The questions were designed to range from easy to difficult. In order, students read a passage, answered corresponding questions, were shown the questions again, and made postdictions. Results showed that students were more aware of their performance for easier items.

Number of test items. Test reliability is a variable that also may influence calibration. Weaver (1990, Exp. 2) manipulated the number of questions on a comprehension test to examine whether adding additional items to the comprehension test would improve calibration accuracy. Weaver asked students to read the expository texts used by Glenberg and Epstein (1985) and to answer inference questions. After reading a text, students made a prediction about their performance, and answered one, two, or four questions. The data showed that students who answered two or four questions per text were able to calibrate their comprehension at greater-than-chance levels (mean G = .28 and .40, respectively), while students who answered only one question per text did no better than chance (mean G = .02). Gamma correlations were significantly higher for the four-question group than for the two-question and one-question groups, but the latter two did not differ from each other. Weaver concluded that reports of low calibration accuracy in previous studies were the result of using only one item per text to measure comprehension. Although increasing the number of items on a comprehension test increases test reliability, the number of test items cannot explain low levels of calibration accuracy that have been reported in the literature (Morris, 1995). Findings of poor calibration have been reported by researchers who have included as many as 40-60 items on comprehension tests (e. g., Magliano, Little, & Graesser,1993; Maki, Foley, Kajer, Thompson, & Willert, 1990; Maki, Jonas, & Kallod, 1994).

Reliability of calibration measure. It is also unlikely that reports of poor calibration are the result of different, perhaps unreliable, measures. Lin, Moore, and Zabrucky (2001) found that students showed consistent calibration accuracy across a variety of measures. In this study, students completed self-reports about their general monitoring ability and took the Nelson-Denny reading comprehension test. They then

read 12 short expository texts and rated how well they could answer inference questions (predictions), how well they understood the text, how easy they found the text, and how interesting they found the text. After this, they answered 48 true/false items, made postdictions (a qualitative measure), and reported the number of items they thought they answered correctly (a quantitative measure). Lin et al. reported that students were able to make both accurate predictions and postdictions. Although two of the pretest scales focused on subject variables (how much one understands a text, or how confident one is), and two focused on text variables (how interesting a text was or how easy a text was to read), the different measures of calibration were highly correlated with one another. In other words, students' calibration accuracy remained stable across different scales. These findings are noteworthy since researchers have used a variety of measures to assess calibration without directly comparing performance across measures to determine if the scales were equally effective measures.

Time delay. In Nelson and Narens' (1990, 1994) model of metamemory, judgments of learning (*JOL*s) are predictions about future test performance that can be made during or after the learning process. A consistent finding in the metamemory literature is that JOLs made immediately after learning are not as accurate as JOLs that are made after a short delay. Dunlosky and Nelson (1994) referred to this phenomenon as the delayed JOL effect and hypothesized that immediate JOLs are less accurate because they are based on information still in short-term memory. However, the recently learned information may not enter long-term memory or, if it does, may not be retrievable after a delay. When there is a time lapse between learning and the JOL judgment, it will be more accurate because participants are basing their JOLs on information in long-term

memory, which is more indicative of their future performance (Dunlosky & Nelson, 1994, 1997).

Glenberg and Epstein (1985, Exp. 1) proposed that, because readers generally make confidence ratings immediately after reading, they may base their confidence ratings on information in short-term memory resulting in inaccurate predictions. They tested this hypothesis by varying the time between reading and predictions. Students read 15 one-paragraph expository texts on a variety of topics and made a prediction after each text (immediate condition) or after reading all 15 texts (delay condition). Results showed that the time delay had no effect on prediction accuracy. Nevertheless, it may be premature to abandon the hypothesis that a time delay may improve calibration accuracy. Because students in the delay condition read all 15 texts and then made predictions, the time delay may have been confounded with the increased task demands of reading additional passages. Students' prediction accuracy may have, in fact, benefited from the time delay, but this was not apparent because students were required to turn their attention to comprehending new passages. Further research is needed to obtain a clear picture of the effects of time.

According to Maki (1998), predicting comprehension performance requires an assessment not only of how well the text was understood, but also how much forgetting will occur. Therefore, adults may adjust their performance assessments by recognizing that information is harder to recall after a delay, or that particular texts may be less memorable than others. To determine whether adults factor forgetting into their prediction ratings, Rawson, Dunlosky, and McDonald (2002) gave students either an open-book test or a closed-book test. The texts were 13 two-sentence expository passages. Students made a calibration judgment after each passage and, after reading all

of the passages, took one of the two tests. Comprehension performance did not vary as a result of test form. The authors explained that students taking the open-book test should not have had an advantage because correct responses depended on text comprehension rather than memory. Looking at the calibration judgments, the authors found that students who took the closed-book test gave lower prediction ratings, suggesting that they were allowing for some forgetting across texts. In Experiment 2, the authors used the same materials but varied the time between reading and the comprehension test. Students were told that they would be tested on half of the texts 15 minutes later or two weeks later. Students made either a prediction rating or a comprehension rating. Although comprehension ratings did not differ due to the anticipated delay, the prediction ratings were significantly lower when students anticipated a two-week delay. These findings lend additional support to the authors' hypothesis that students factor forgetting into their prediction ratings.

Type of calibration judgment. The literature suggests that postdictions are generally more accurate than predictions. One possible explanation for this is that adults rely on different information when making predictions and postdictions. The metamemory literature provides strong evidence that EOL, JOL, and FOK judgments are not highly correlated, indicating that the judgments monitor different aspects of memory (Nelson & Narens, 1994). This raises the possibility that predictions (which would most closely resemble EOL judgments) and postdictions (which would be most similar to JOL judgments) tap different aspects of metacomprehension. There is some evidence to suggest that this is the case. First, a consistent finding in the literature is that readers are able to make accurate postdictions even when their predictions are no better than chance (Pierce & Smith, 2001). Second, variables such as reading ability (Maki, Jonas, &

Kallod, 1994), rereading (Rawson, Dunlosky, & Thiede, 2000), text difficulty (Maki et al., 1990), and prior knowledge (Glenberg & Epstein, 1987) have been found to influence prediction accuracy but not postdiction accuracy.

Only one study has directly compared prediction and postdiction accuracy. According to Pierce and Smith (2001), retrieval hypotheses propose that confidence judgments are based on what students remember about their performance on a particular test. As a result, postdictions will be consistently more accurate than predictions because readers cannot gain feedback from a particular test until after they have taken it. Alternatively, readers may use their prior knowledge about the nature of tests when making postdictions (test knowledge hypotheses). If this is the case, readers should be able to predict their future test performance as accurately as their past test performance once they have acquired sufficient general knowledge about test taking.

To test these hypotheses, Pierce and Smith (2001) had students read four narrative and three expository texts. After reading each text, students answered 16 test items, making predictions and postdictions for each set of four items. Results showed that students were much better at postdictions (mean G = .47) than predictions (mean G = .05) when assessing their performance on the narrative texts. All four postdiction correlations were significant compared to only one of the prediction correlations. Similarly, postdictions (mean G = .53) were significantly higher than predictions (mean G = .28) for the expository texts. For all three texts, only the postdiction correlations were significant. The finding that students' prediction accuracy did not improve from one set of items to the next indicates that students were not able to use their knowledge about the first set of test items to improve their prediction accuracy on the subsequent sets (supporting retrieval hypotheses). Second, students' postdictions were consistently more accurate

than their predictions, suggesting that students may base their predictions and postdictions on different aspects of metacomprehension.

Text Variables

Text difficulty. Previous research has been inconclusive about the influence of information type on calibration accuracy. Results of several studies were mixed as to whether students display greater calibration accuracy for questions that require knowledge of main points/themes or details. Weaver and Bryant (1995, Exp. 1) proposed that whether students attend to the theme of a passage or to the details may depend on what type of texts they are reading. They asked students to read either four narrative or four expository texts. After reading a text, students were asked to predict their future test performance on a test and, after reading all four texts, were given a surprise comprehension test. The data showed that students who read narratives were significantly more accurate when predicting their performance on thematic questions than on detailed questions, while students who read expository texts showed the reverse pattern.

Because the narrative passages had easier readability levels than the expository passages, Weaver and Bryant (1995) conducted a second experiment in which they controlled for text difficulty and found that the interaction between text type and information type disappeared. In Experiment 2, the authors used Flesch's (1951) scale to determine readability levels of texts. Students read texts that were either *easy* (below grade 8), *standard* (approximately grade 12), or *difficult* (approximately grade 16). The authors found that both test performance and prediction accuracy were influenced by the readability levels. As might be expected, students who read easy passages did significantly better on the comprehension test (61% correct) than those who read standard

(49% correct) or difficult (46% correct) texts. However, students who read the standard texts were significantly more accurate when predicting their performance than students who read easy or difficult passages. Why was calibration improved when students read standard texts? Weaver and Bryant proposed that readers did not need to utilize any metacognitive processes when comprehending easy texts. Conversely, difficult texts required that readers devote all of their cognitive processing to text comprehension, leaving few resources available for metacognitive judgments. The authors suggested that readers should display the greatest calibration accuracy for reading tasks that required an intermediate level of cognitive effort (the *optimum effort hypothesis*). Because many calibration studies have been conducted using texts developed by Glenberg and Epstein (1985), Weaver and Bryant conducted readability analyses on those texts and found that all but one scored in the difficult range. They concluded that the difficulty level of the texts explained the low levels of calibration accuracy reported in the literature.

Subsequent studies have failed to support the optimum effort hypothesis. Pierce and Smith (2001) used Weaver and Bryant's (1995) texts and found that students were able to make accurate predictions only when questions related to difficult texts. In Experiment 1, students read narrative texts (two easy, one standard, and one difficult). According to the optimum effort hypothesis, the standard texts should have produced the highest calibration accuracy. However, results revealed there was no difference in postdiction accuracy across the four texts, and the only prediction correlation that was significant was for the difficult text (G = -.44). In Experiment 2, students read expository texts (two rated standard and one easy). Students were no better at predicting their performance on the standard texts than on the easy text.

Palmer (1995) was also unable to support the optimum effort hypothesis when comparing the prediction accuracy of college freshmen and seniors. Students read an *easy* passage (Fry readability grade level 7) or a *difficult* passage (Fry readability grade level 12), after which they rated their understanding of the text and took a comprehension test. According to the optimum effort hypothesis, students should have shown the higher level of prediction accuracy for the difficult passage; however, freshmen were poor at predicting their performance regardless of passage difficulty, and seniors were better at predicting their performance for the easy passage.

In a study that investigated whether age influences adults' monitoring ability, Lin, Zabrucky, and Moore (2002) also were unable to support the optimum effort hypothesis. They used the same texts as Weaver and Bryant (1995) and followed the same design. Participants also completed the vocabulary and comprehension subscales of the Nelson-Denny Reading Test (Form E) to determine reading levels, which indicated that younger adults read at an average level of grade 16.4 and older adults at a level of grade 13.7. Using the Flesch (1951) readability scale, the texts were designated as *easy* (below grade 8), *standard* (approximately grade 12), or *difficult* (approximately grade 16). Results showed that participants were more accurate at calibration for standard and easy texts (G = .47 and .19, respectively) than for difficult texts (G = -.09). According to the optimal effort hypothesis, participants should have shown the highest levels of calibration accuracy for the difficult passages, when, in fact, prediction accuracy was poorest for those passages.

Improving Calibration Accuracy

So far we have examined the influence of a number of variables on calibration accuracy. While the variables affect calibration to different degrees, a consistent finding

across all of the studies is that adults are poor at calibrating their performance. Thus, several researchers have investigated possible ways of improving calibration accuracy.

Rereading. Results from error detection studies have shown that rereading improved error detection rates in some cases. Rawson et al. (2000) found that rereading also improves calibration accuracy. In Experiment 1, students read six expository texts, made predictions, and took a comprehension test. The texts were presented one sentence at a time on a computer screen, and students made predictions after reading each text. The rereading group read each text twice before making predictions; a control group read the texts one time. Although rereading did not have an effect on comprehension performance, students who reread texts were significantly more accurate in their predictions (G = .57) than the control group (G = .24). In Experiment 2, each text was displayed in its entirety on the computer screen rather than displayed one sentence at a time, and students made predictions after reading all six texts rather than after reading each text. This was done to determine if a time delay would have an effect on the rereading manipulation. Contrary to the results of Experiment 1, results showed a significant difference on comprehension performance between the rereading group and the control group (86% correct and 70% correct, respectively). The time delay, however, did not affect prediction accuracy. Similar to Experiment 1, students who reread the texts were significantly more accurate in their predictions than the control group (G = .55 and .19, respectively). Although not discussed by the authors, it is interesting that rereading had an effect on comprehension performance in Experiment 2 but not in Experiment 1. One possible explanation is that, because students read all six texts before making predictions in Experiment 2, rereading lowered interpassage confusion which improved comprehension performance.

Dunlosky and Rawson (2005) also found that rereading improved students' prediction accuracy. Students were assigned to one of three groups: single reading, immediate rereading, or delayed rereading. The authors included a delay condition to determine if the benefits of rereading found in previous studies would persist over time. The materials were six short expository texts that were presented one sentence at a time on a computer screen. In the single reading condition, students read each text once, made a prediction, and completed a comprehension test. In the immediate rereading condition, students read all of the texts one time, and then reread them before making a prediction and taking the comprehension test. The delayed rereading group read the texts once, returned one week later, read each text again, made a prediction, and completed the comprehension test. Dunlosky and Rawson found that students who reread the texts immediately were significantly better able to predict their performance than those who were permitted to read the texts only one time. Being given the opportunity to reread after a one-week delay, however, had no effect on prediction accuracy. Students in the delayed rereading group performed at about the same level as those who read the texts only once. Recall that Rawson et al. (2000) reported that rereading improved performance on the comprehension test in one experiment, but not the other. Results from the present study showed that rereading had no effect on comprehension test performance.

Practice. Several studies have investigated whether giving students practice on calibration tasks will improve their performance, with mixed results. In Glenberg and Epstein's (1985, Exp. 2) study, half of the students practiced reading texts and answering comprehension questions. All the students then read 15 texts, made predictions, and answered one true/false question per text. There was no effect of practice on test

performance or prediction accuracy. In Experiment 3, however, the authors found that students could benefit from practice. In order, students made a prediction, answered a comprehension question, made a postdiction, made a second prediction, and answered a second question. The authors found that students' prediction accuracy significantly improved from the first question to the second question. However, in a later experiment, Glenberg and Epstein (1987) used the same design and found that students were unable to use their experience with the first question to improve their calibration on the second question.

In an attempt to resolve these inconsistencies, Glenberg et al. (1987, Exp. 6) varied the type of question on the practice test and found that practice was helpful, but only when practice questions were similar to those on the actual test. They asked students to read 16 passages and then take a practice test, after which they predicted their performance on a subsequent test over the same texts. The practice test items were either the same (a problem identical to the actual test), related (a paraphrase of an identical problem) or unrelated (different from the actual test). Only those students who answered practice questions that were identical to those on the actual test showed improvement in prediction accuracy. Similar results were obtained in three additional experiments. The authors concluded that practice in the form of sample questions could improve prediction accuracy only if the sample questions were very similar to those on the actual test.

Maki and Serra (1992b) varied the task demands and found that students did not benefit from practice. Rather than predict their comprehension performance after reading, students rank ordered texts according to comprehension difficulty. They then answered comprehension questions and rank ordered the texts a second time. This procedure was repeated for three sets of four texts. Results showed that practice had no

effect on ranking accuracy. Maki and Serra's contradictory findings may be the result of a difference in task demands (rank ordering texts rather than answering direct questions about comprehension test performance). Another possibility is that students did not base their predictions of test performance on the comprehension difficulty of the texts, but relied instead on other variables (e.g., their reading ability, prior knowledge, confidence bias).

Feedback. In the studies discussed above, practice had little effect on students' calibration accuracy unless practice tests were identical to criterion tests. Students may have benefited from the practice tests if they had been given feedback on their performance. In Maki and Berry's (1984) study of the effects of reading ability on calibration, some of the students were given feedback about their performance. Recall that students made predictions after reading sections of an introductory psychology text and took a comprehension test the next day. After the first test, half of the students were told how well they had performed on the test. The procedure (reading, making predictions, and taking a comprehension test the following day) was repeated. The authors found that giving students feedback did not improve their prediction accuracy on the second test.

Schraw, Potenza, and Nebelsick-Gullet (1993) also concluded that feedback had little effect on calibration accuracy. Students completed the Nelson Denny comprehension test (Form E, 1981) and answered 36 multiple-choice questions divided into 8 subtests of 4-8 questions each. Students made a postdiction after each question. Half were given feedback about their performance after completing each subtest; however, their postdiction accuracy showed no improvement on subsequent subtests.

Maki and Serra (1992b, Exp. 3) asked students to read four texts and take a practice test. Half of the students were given feedback about their performance. While feedback had no effect on postdiction accuracy, it actually reduced students' prediction accuracy. When the authors subsequently compared scores on the practice test and the criterion test, they found that the correlation between the two was quite low. They concluded that giving students feedback on the practice test may have reduced prediction accuracy because their performance on the practice test bore little resemblance to their performance on the criterion test.

Rather than feedback about test performance, Walczyk and Hall (1989) gave students an opportunity to give themselves feedback about their level of text comprehension as they read a chapter on descriptive statistics. Students read one of four versions: text only, text with examples, text with questions after each paragraph, or text with both examples and questions. Walczyk and Hall reasoned that inserting examples and questions in the text would give students the opportunity to test themselves about the chapter material and that self testing would have a beneficial effect on both test performance and postdiction accuracy. Although the different levels of self-testing had no effect on comprehension performance, students' postdictions were significantly more accurate when they read the chapter that contained both examples and questions compared to the other three groups.

Levels of processing. The embedded examples and questions in Walczyk and Hall's (1989) study may have had the effect of increasing students' processing of the text which, in turn, improved their prediction accuracy. Several researchers have used manipulations that increased levels of text processing and in some, but not all cases, reported greater calibration accuracy. To increase text processing demands, Maki et al.

(1990) altered texts so that some paragraphs contained words with deleted letters. After reading a paragraph, students rated the comprehension ease of the paragraph or made a prediction rating, then answered two cued recall questions, and made a postdiction. The authors found that the increased processing required to comprehend deleted-letters paragraphs had the effect of improving both comprehension performance and prediction accuracy.

Magliano et al. (1993), however, found that asking students to use either shallow or deep processing had no effect on either performance or prediction accuracy. Students listened to taped instructions of either deep or superficial reading strategies. The deep reading strategies were instructions to clarify unclear concepts, make predictions about upcoming material, form relevant questions about the text, and summarize the text. The superficial strategies focused on basic linguistic skills, local decoding skills, and motivation. After listening to the instructions, students read expository texts, made predictions and took a comprehension test. Type of instructions did not affect either performance or prediction accuracy. In a second experiment, the authors added a control group that received no reading strategies. Again, the groups did not differ on any of the measures.

Instead of giving specific strategy instructions to promote shallow or deep processing, Schommer and Surber (1986) asked students to read a passage and rate its clarity (shallow processing instructions) or read a passage with the goal of teaching another student its contents (deep processing instructions). After reading, students answered three multiple choice questions and made a postdiction. Students who were given deep processing instructions were more accurate at calibrating their performance than those who received shallow processing instructions.

Incentives. Rather than give students instructions to increase text processing, Schraw et al. (1993) examined whether giving students rewards would increase text processing and calibration accuracy. They assigned students to one of three incentive conditions: (1) no incentive, (2) double credit if performance scores were above the group mean, and (3) double credit if calibration was within 25% of perfect. Students took a mathematical probability test and the Nelson Denny Reading Test (Form E), making postdictions after each test. Schraw et al. found that offering students a reward for performance had no effect on either performance or monitoring accuracy but that rewarding them for improved calibration improved performance rather than calibration accuracy. The authors hypothesized that giving incentives for calibration accuracy may have encouraged students to think about reasons why they performed well in general rather than think about how well they performed on the experimental tests.

In sum, there appears to be a connection between the processing demands of a comprehension task and calibration performance. Increasing the likelihood that readers will process text at a deeper level has been an effective way of improving calibration accuracy in some studies but not others.

An important part of learning is being able to make accurate decisions about when studied material has been mastered (Maki, 1998). Unfortunately, the evidence from calibration studies shows that adults are often poor at evaluating their comprehension. This is of particular concern because the vast majority of participants in the calibration studies reviewed in this paper have been college undergraduates.

The evidence also shows that evaluation accuracy is influenced by a number of subject, task, and text variables. Findings for subject variables suggest that calibration may be influenced by a number of factors. While there is little support for the hypothesis

that adults develop a general metacognitive ability based on metacognitive knowledge, students often use their domain knowledge as well as information in a text when assessing their comprehension. Findings also suggest that adults use their prior performance, and their beliefs about learning and themselves as learners to predict their future performance. Furthermore, some adults may possess a stable trait of confidence bias that mediates their calibration judgments.

With respect to task variables, the most consistent finding is that adults are better at postdictions than predictions. Postdictions are likely to be less demanding than predictions because readers have additional information about the nature of the test as well as their performance when making postdictions. Elements of the comprehension test (e.g., item difficulty, item format) have not been extensively investigated. When task demands require students to use deep processing, students show improvement on calibration tasks. Little is known about the influence of text variables on calibration. Text difficulty may be a factor, but not in all cases.

Conclusion

What are the main conclusions that emerge from this body of literature? Looking at stable metacomprehension knowledge, it is not surprising to find that adults know about and use a variety of strategies to accomplish their comprehension goals. It is also not surprising that knowledge about strategies and tasks is related to both comprehension performance and academic achievement.

Perhaps more surprising is that knowledge about person variables may play a major role in comprehension performance. When Flavell (1979) first described metacognitive knowledge, he referred to the relevant factors as “cognitive variables.” He later proposed that a more accurate definition would be “the part of one’s acquired world

knowledge that has to do with cognitive (or perhaps better, psychological) matters" (Flavell, 1987, p. 21). Only recently have researchers begun to examine how psychological variables such as motivation, beliefs, and self-assessments of ability influence comprehension performance. Questionnaires such as the MCS (Moore, Zabrucky, Commander, & Morton, 1993) and Schommer's (1990) epistemological beliefs questionnaire have shown that several psychological variables are predictors of comprehension performance. For example, adults who feel confident about their comprehension abilities and believe they have control over their comprehension tend to do better on comprehension tasks, whereas adults who report stress related to comprehension activities tend to do poorly. Future research may show that these psychological factors are a critical part of successful comprehension monitoring.

Looking at regulation of comprehension, results from error detection studies show that adults are far below ceiling at regulating their understanding. However, it may be difficult to accurately determine adults' regulation ability because of the methodological problems associated with the research paradigms that have been used. Perhaps the largest obstacle is that reading has become such an automatic process for adults that they are unable to accurately report on their regulation activities. The data that adults alter their eye movements when they encounter errors in text, yet fail to report the errors, suggests that many errors in texts are recognized but not brought to conscious awareness. Brown (1987) used the term *reflective access* to describe the ability to talk about cognitive functions, and noted that knowing how to do something did not necessarily mean that one could describe what he or she was doing. Thus, it may be that some metacomprehension activities are simply not statable.

The evidence from calibration studies shows that adults are generally poor at accurately assessing their level of text comprehension, but are better at judging their past performance on comprehension tasks than they are at judging their future performance. A common finding is that many adults have illusion of knowing, and are often overconfident about their comprehension performance. An interesting hypothesis is that some adults develop a bias about themselves as learners that interferes with their ability to accurately monitor their performance on cognitive tasks. These adults tend to be consistently overconfident (or underconfident) regardless of their actual performance.

What we have learned from both knowledge of comprehension and monitoring of comprehension has given us some insight into differences between good and poor readers. Although poorer readers are able to monitor their comprehension, they evaluate their comprehension less frequently than better readers and use less sophisticated strategies to resolve comprehension failures. It is not clear why this is the case. One answer is that poorer readers do not have the requisite knowledge of task and strategy variables that permits higher levels of text comprehension. Alternatively, poorer readers may have the knowledge, but don't apply it. As Brown (1980) points out, readers' goals vary and the comprehension strategies they use must be judged according to the goals they set. Poorer readers may not use more sophisticated strategies because they approach comprehension tasks with the goal of remembering rather than learning. Whereas better readers attempt to use information gained from text to help solve problems or integrate new information with prior knowledge, poorer readers may focus on recalling information for a test. Also, poorer readers may have beliefs that interfere with successful comprehension monitoring. They may, for example, believe that they lack

good comprehension skills and cannot improve them. Future research should explore how person variables such as motivation and beliefs influence poorer readers.

Additional research is also needed with respect to how task and text variables affect comprehension monitoring. In particular, future research needs to clarify the role of rereading. Rereading is perhaps the most commonly used compensatory strategy when adults encounter comprehension failures. It is also a successful strategy. When adults reread, they generally do better on both error detection tasks and calibration tasks. Somewhat confusing, however, is that results show that rereading does not always improve actual comprehension performance.

A second source of confusion is that reading ability appears to have little effect on evaluation of comprehension. Results from both self-reports and error detection studies show that reading ability is related to academic achievement, more frequent regulation of comprehension, and the use of more sophisticated strategies. However, results from calibration studies indicate that adults with stronger verbal skills are generally no better at calibrating their performance than adults with weaker skills. Future studies should address the question of whether this is an accurate finding or a methodological artifact.

References

Afflerbach, P., & Johnston, P. (1984). Research methodology on the use of verbal reports in reading research. *Journal of Reading Behavior, 16,* 307-322.

Alessi, S. M., Anderson, T. H., & Goetz, E. T. (1979). An investigation of lookbacks during studying. *Discourse Processes, 2,* 197-212.

Baker, L. (1979). Comprehension monitoring: Identifying and coping with text confusions. *Journal of Reading Behavior, 11,* 365-374.

Baker, L. (1985). Differences in the standards used by college students to evaluate their comprehension of expository prose. *Reading Research Quarterly, 20,* 297-313.

Baker, L., & Anderson, R. I. (1982). Effects of inconsistent information on text processing: Evidence for comprehension monitoring. *Reading Research Quarterly, 17,* 281-294.

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. D. Pearson (Ed.), *Handbook of research in reading* (pp. 353-395). New York: Longman.

Brennan, S., Winograd, P. N., Bridge, C. A., & Hiebert, E. H. (1986). A comparison of observer reports and self-reports of study practices used by college students. In J. A. Niles & R. V. Lalik (Eds.), *Solving problems in literacy: Learners, teachers, and researchers: Thirty-fifth yearbook of the National Reading Conference* (pp. 353-358). Rochester, NY: National Reading Conference.

Brown, A. (1980). Metacognitive development and reading. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 453-481). Hillsdale, NJ: Erlbaum.

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65-116). Hillsdale, NJ: Erlbaum.

Cioffi, G. (1986). Relationships among comprehension strategies reported by college students. *Reading Research and Instruction, 25,* 220-231.

Commander, N. E., & Stanwyck, D. J. (1997). Illusion of knowing in adult readers: Effects of reading skill and passage length. *Contemporary Educational Psychology, 22,* 39-52.

De Carvalho Filho, M. K., & Yuzawa, M. (2001). The effects of social cues on confidence judgments mediated by knowledge and regulation of cognition. *Journal of Experimental Education, 69,* 325-345.

Dunlosky, J., & Hertzog, C. (2001). Measuring strategy production during associative learning: The relative utility of concurrent versus retrospective reports. *Memory & Cognition, 29,* 247-253.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of Judgments of Learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33,* 545-565.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34,* 906-911.

Flavell, J. H. (1981). Cognitive monitoring. In W. P. Dickson (Ed.), *Children's oral communication skills* (pp. 35-60). New York: Academic Press.

Flavell, J. H. (1987). Speculations about the nature and development of metacognition. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 21-29). Hillsdale, NJ: Erlbaum.

Garner, R. (1981). Monitoring of passage inconsistency among poor comprehenders: A preliminary test of the "piecemeal processing" explanation. *Journal of Educational Research, 74,* 159-162.

Garner, R. (1982). Verbal-report data on reading strategies. *Journal of Reading Behavior, 14,* 159-167.

Garner, R., & Alexander, P. A. (1989). Metacognition: Answered and unanswered questions. *Educational Psychologist, 24,* 143-158.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 702-718.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15,* 84-93.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116,* 119-136.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition, 10,* 597-602.

Grabe, M., Antes, J., Kahn, H., & Kristjanson, A. (1991). Adult and adolescent readers' comprehension monitoring performance: An investigation of monitoring accuracy and related eye movements. *Contemporary Educational Psychology, 16,* 45-60.

Grabe, M., Antes, J., Thorson, I., & Kahn, H. (1987). Eye fixation patterns during informed and uninformed comprehension monitoring. *Journal of Reading Behavior, 19,* 123-140.

Hacker, D. J. (1998). Self-regulated comprehension during normal reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 165-191). Mahwah, NJ: Erlbaum.

Hare, V. C. (1981). Readers' problem identification and problem solving strategies for high- and low-knowledge articles. *Journal of Reading Behavior, 13,* 359-365.

Kambe, G., Rayner, K., & Duffy, S. A. (2001). Global context effects on processing lexically ambiguous words: Evidence from eye fixations. *Memory & Cognition, 29,* 363-372.

Karabenick, S. A. (1996). Social influences on metacognition: Effects of colearner questioning on comprehension monitoring. *Journal of Educational Psychology, 88,* 689-701.

Kaufman, N. J., Randlett, A. L., & Price, J. (1985). Awareness of the use of comprehension strategies in good and poor college students. *Reading Psychology: An International Quarterly, 6,* 1-11.

Keleman, W. L., Frost, P. J., & Weaver, C. A. III (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition, 28,* 92-107.

Kroll, M. D., & Ford, M. L. (1992). The illusion of knowing, error detection, and motivational orientations. *Contemporary Educational Psychology, 17,* 371-378.

Lin, L., Moore, D., & Zabrucky, K. M. (2000). Metacomprehension knowledge and comprehension of expository and narrative texts among younger and older adults. *Educational Gerontology, 26,* 737-749.

Lin, L., Moore, D., & Zabrucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology, 22,* 111-128.

Lin, L., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23,* 345-391.

Lin, L., Zabrucky, K. M., & Moore, D. (1997). The relations among interest, self-assessed comprehension, and comprehension performance in young adults. *Reading Research and Instruction, 36,* 127-139.

Lin, L., Zabrucky, K. M., & Moore, D. (2002). Effects of text difficulty and adults' age on relative calibration of comprehension. *American Journal of Psychology, 115,* 187-198.

Lorch, R. F., & van den Broek, P. (1997). Understanding reading comprehension: Current and future contributions of cognitive science. *Contemporary Educational Psychology, 22,* 213-246.

Lundeberg, M. A. (1987). Metacognitive aspects of reading comprehension: Studying understanding in legal case analysis. *Reading Research Quarterly, 22*, 407-432.

Lundeberg, M. A., Fox, P. W., & Puncochar, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology, 86,* 114-121.

Magliano, J. P., Little, L. D., & Graesser, A. C. (1993). The impact of comprehension instruction on the calibration of comprehension. *Reading Research and Instruction, 32,* 49-63.

Maki, R. H. (1995). Accuracy of metacomprehension judgments for questions of varying importance level. *American Journal of Psychology, 108,* 327-344.

Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117-144). Mahwah, NJ: Erlbaum.

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., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 609-616.

Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin and Review, 1,* 126-129.

Maki, R. H., & McGuire, M. J. (2002). Metacognition for text: Findings and implications for education. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 39-67). Cambridge, UK: Cambridge University Press.

Maki, R. H., & Serra, M. (1992a). The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 116-126.

Maki, R. H., & Serra, M. (1992b). Role of practice tests in the accuracy of test predictions on text material. *Journal of Educational Psychology, 84,* 200-210.

Moore, D., Lin-Agler, L. M., & Zabrucky, K. M. (2005). A source of metacomprehension inaccuracy. *Reading Psychology, 26,* 251-265.

Moore, D., Zabrucky, K., & Commander, N. E. (1997a). Validation of the metacomprehension scale. *Contemporary Educational Psychology, 22*, 457-471.

Moore, D., Zabrucky, K., & Commander, N. E. (1997b). Metacomprehension and comprehension performance in younger and older adults. *Educational Gerontology, 23*, 467-475.

Moore, D., Zabrucky, K., Commander, N. E., & Morton, J. L. (1993, March). *Validity of the metacomprehension questionnaire*. Paper presented at the meeting of the Southeastern Psychological Association, Atlanta, GA.

Morris, C. C. (1995). Poor discourse comprehension monitoring is no methodological artifact. *The Psychological Record, 45,* 655-668.

Myers, S. S. (1991). Performance in reading comprehension--product or process? *Educational Review, 43,* 257-272.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109-133.

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. *Applied Cognitive Psychology, 10,* 257-260.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 125-173). New York: Academic Press.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. A. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1-25). Cambridge, MA: MIT Press.

Pallier, G. (2003). Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles, 48,* 265-276.

Pallier, G., Wilkinson, R., Danthir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology, 129,* 257-299.

Palmer, B. M. (1995). Comprehension calibration of college freshmen and college seniors. In K. A. Hinchman, D. J. Leu, & C. K. Kinzer (Eds.), *Perspectives on literacy research and practice: Forty-fourth yearbook of the National Reading Conference* (pp. 205-211). Chicago: National Reading Conference.

Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition, 29,* 62-67.

Pressley, M., & Ghatala, E. S. (1988). Delusions about performance on multiple-choice comprehension tests. *Reading Research Quarterly, 23,* 454-464.

Pressley, M., Ghatala, E. S., Woloshyn, V., & Pirie, J. (1990). Sometimes adults miss the main ideas and do not realize it: Confidence in responses to short-answer and multiple-choice comprehension questions. *Reading Research Quarterly, 25,* 233-249.

Raney, G. E., & Rayner, K. (1991). Event-related brain potentials, eye movements, and reading. *Psychological Science, 4,* 283-286.

Rawson, K. A., Dunlosky, J., & McDonald, S. L. (2002). Influences of metamemory on performance predictions for text. *The Quarterly Journal of Experimental Psychology, 55A,* 505-524.

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28,* 1004-1010.

Rayner, K. (1993). Eye movements in reading: Recent developments. *Current Directions in Psychological Science, 2,* 81-85.

Schiefele, U. (1992). Topic interest and levels of text comprehension. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 151-182). Hillsdale, NJ: Erlbaum.

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology, 82,* 498-504.

Schommer, M., Crouse, A., & Rhodes, N. (1992). Epistemological beliefs and mathematical text comprehension: Believing it is simply does not make it so. *Journal of Educational Psychology, 84,* 435-443.

Schommer, M., & Surber, J. R. (1986). Comprehension-monitoring failure in skilled adult readers. *Journal of Educational Psychology, 78,* 353-357.

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology, 19,* 143-154.

Schraw, G. (1997). The effect of generalized metacognitive knowledge on test performance and confidence judgments. *The Journal of Experimental Education, 65,* 135-146.

Schraw, G. (1998). On the development of adult metacognition. In M. C. Smith & T. Pourchot (Eds.), *Adult learning and development* (pp. 89-106). Mahwah, NJ: Erlbaum.

Schraw, G., Dunkle, M. E., Bendixen, L. D., & Roedel, T. D. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology, 87*, 433-444.

Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology, 18*, 455-463.

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures* (2nd ed.). New York: Chapman & Hall.

Spring, C. (1985). Comprehension and study strategies reported by university freshmen who are good and poor readers. *Instructional Science, 14*, 157-167.

Taraban, R., Rynearson, K., & Kerr, M. S. (2000). Metacognition and freshman academic performance. *Journal of Developmental Education, 24*, 12-26.

Walczyk, J. J., & Hall, V. C. (1989). Effects of examples and embedded questions on the accuracy of comprehension self-assessments. *Journal of Educational Psychology, 81*, 435-437.

Walczyk, J. J., Marsiglia, C. S., Bryan, K. S., & Naquin, P. J. (2001). Overcoming inefficient reading skills. *Journal of Educational Psychology, 93*, 750-757.

Weaver, C. A. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 214-222.

Weaver, C. A., & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition, 23*, 12-22.

Winograd, P., & Johnston, P. (1982). Comprehension monitoring and the error detection paradigm. *Journal of Reading Behavior, 14*, 61-76.

Yussen, S. R., & Smith, M. C. (1990). Detecting general and specific errors in expository texts. *Contemporary Educational Psychology, 15*, 224-240.

Zabrucky, K. (1990). Evaluation of understanding in college students: Effects of text structure and reading proficiency. *Reading Research and Instruction, 29,* 46-54.

Zabrucky, K. M., & Cummings, A. M. (2004). Metacognition. In C. Fisher & R. Lerner (Eds.), *Encyclopedia of applied developmental science* (pp. 723-727). Thousand Oaks, CA: Sage.

Zabrucky, K., & Moore, D. (1994). Contributions of working memory and evaluation and regulation of understanding to adults' recall of texts. *Journal of Gerontology: PSYCHOLOGICAL SCIENCES, 49,* P201-P212.

Zabrucky, K. M., & Moore, D. (1999). Influence of text genre on adults' monitoring of understanding and recall. *Educational Gerontology, 25,* 691-710.

CHAPTER 2

THE USE OF ITEM RESPONSE THEORY TO ASSESS ADULTS' POSTDICTION ACCURACY

The ability to read and comprehend text is crucial for academic success. Because of its importance, a number of researchers have studied the metacognitive aspects of reading comprehension (*metacomprehension*) (Maki, 1998). Metacomprehension can be divided into a relatively permanent, stable component (*knowledge*) and a less stable, on-line component (*monitoring*). The knowledge component includes what a person has learned and believes about comprehension strategies (e.g., rereading aids comprehension), task demands (e.g., a text with many unfamiliar words will be more difficult to comprehend than a text with many familiar words), and subject variables (e.g., reading ability). Monitoring of comprehension is a two-part process, in which readers must keep track of how well comprehension is proceeding (*evaluation*) and initiate actions to resolve any comprehension failures that occur (*regulation*) (Baker, 1979). Evaluation is critical to successful monitoring because readers cannot initiate strategies to fix comprehension failures if they are not aware that such failures have occurred (Zabrucky, 1990). Regulation is as critical as evaluation because awareness is not enough. Readers must be willing and able to deal with comprehension failures after they have been detected (Baker, 1985).

Researchers interested in the evaluation component of comprehension monitoring have often used the calibration paradigm. In the calibration paradigm, readers are asked

to read one or more texts and then make a confidence judgment about their future performance on a test over the material (a *prediction*). Alternatively, readers may be asked to assess their performance after they have taken a comprehension test (a *postdiction*, sometimes referred to as calibration of performance) (Glenberg, Sanocki, Epstein, & Morris, 1987). Calibration is the relation between subjective confidence judgments and actual test performance.

Using the calibration paradigm, researchers have examined a number of variables to determine their possible effects on evaluation (for a review, see Lin & Zabrucky, 1998). Among these are text variables (e.g., text genre, difficulty level), task demands (e.g., type of reading instruction, type of test items), and individual differences (e.g., reading ability, prior knowledge). The most consistent finding from these studies is that adults tend to be poor at evaluation. Although they are often able to evaluate their performance at above chance levels, reported gamma correlations are generally quite low (in the .20 to .30 range). One explanation for these findings is that adults are, in fact, poor at evaluating their comprehension. An equally plausible explanation, however, is that the research methods used to measure evaluation have not provided reliable estimates of adults' ability (Keleman, Frost, & Weaver, 2000; Maki, 1998).

Previous attempts at explaining low calibration accuracy by citing methodological problems have been unsuccessful. For example, Weaver (1990) proposed that findings of low calibration accuracy in early studies was due to the use of a single item on the comprehension test. Weaver asked students to read a passage, and make predictions about their performance on one, two, or four comprehension questions. Students who answered two or four questions per text were significantly better at predicting their performance than students who answered only one question per text. Because a number

of studies had used only one item per text to measure comprehension, Weaver concluded that findings of low calibration accuracy previously reported were the result of poor reliability. The number of items used on previous comprehension tests cannot, however, explain the low levels of calibration accuracy reported in the literature. Researchers have found poor calibration when participants answered as many as 40 - 60 items on a comprehension test (e.g., Magliano, Little, & Graesser, 1993; Maki, Foley, Kajer, Thompson, & Willert, 1990; Maki, Jonas, & Kallod, 1994; Moore, Lin, & Zabrucky, 2005).

A different approach is to use Item Response Theory modeling to improve the quality of a comprehension test. Item Response Theory (*IRT*) is a modern psychometric approach that has a number of advantages compared to Classical Test Theory (*CTT*). Although the development of IRT has taken place primarily in the area of educational assessment, in recent years researchers have successfully applied IRT in the areas of psychological and educational research (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Junker & Sijtsma, 2001; Reise, Smith, & Furr, 2001; Rouse, Finger, & Butcher, 1999; Silver, Smith, & Greene, 2001; Stark, Chernyshenko, Lancaster, Drasgow, & Fitzgerald, 2002). Because of its psychometric properties, an IRT-based comprehension test may be beneficial to metacomprehension research. In particular, an IRT-developed comprehension test may help explain mixed findings that have been reported with respect to the effects of reading ability and test difficulty on calibration accuracy.

This paper is organized into three parts. First, research findings related to reading ability and test difficulty will be summarized. Then the benefits of IRT will be briefly

discussed. In addition, a rationale for examining the effects of guessing on postdiction accuracy will be offered.

Reading Ability

It seems intuitively obvious that people with increased skill or knowledge in a domain would make more accurate metacognitive judgments (Schwartz & Metcalfe, 1994). However, findings from studies that have investigated the relationship between reading ability and calibration judgments have generally shown that reading ability has no effect on calibration accuracy.

In the only study that found a relationship, Maki and Berry (1984) asked 30 students to read a half-chapter of an introductory psychology text, after which they predicted their performance on a comprehension test over the material. Students returned the next day and answered multiple choice questions on the half-chapter they had read the previous day. This procedure (reading a half-chapter, making one or more predictions, and returning the next day for the comprehension test) was repeated. Maki and Berry used a median split on the comprehension test to classify students as good or poor readers and found that students who scored above the median gave higher confidence ratings for correct than incorrect test items. However, in a second experiment, students were permitted to answer the comprehension test questions immediately after making predictions, and those who scored below the median on the comprehension test were as accurate at calibrating their performance as those who scored above the median.

The majority of subsequent studies also have failed to find that reading ability influenced calibration performance. Pressley and Ghatala (1988) used SAT verbal scores as an independent measure of reading ability and found no correlation between reading ability and postdiction accuracy when they compared students' performance on three

types of reading measures (opposites, analogies, and reading comprehension). Commander and Stanwyck (1997) were also unable to find a relationship between reading ability and calibration accuracy. Students were classified as good or poor readers based on a median split of their Nelson Denny Reading Test (Form C) scores. Students read short expository passages taken from introductory college texts, made predictions, answered comprehension questions, and then made postdictions. The authors combined the prediction and postdiction ratings and found that reading ability had no effect on students' calibration accuracy.

Maki et al. (1994) determined students' reading ability by scores on the Multi-Media Comprehension Battery (MMCB), which assesses comprehension of written, auditory, and pictorial stories, and the Nelson-Denny Reading Test (Form C) which assesses comprehension and reading rate. Students read one of two texts taken from *Science News*, each of which was broken down into 12 sections (4-14 sentences each), made predictions, answered comprehension questions, and made postdictions. Students' prediction accuracy was low, although greater than chance (G = .114). The postdiction judgments were significantly higher than the predictions and greater than zero (G = .551). The only score that correlated significantly with prediction accuracy was the auditory comprehension portion of the MMCB. However, postdiction ratings correlated with both portions of Nelson-Denny and with the written and auditory comprehension portions of the MMCB. Reading ability did not correlate with prediction accuracy, but was related to postdiction accuracy. Better and faster readers made more accurate postdiction judgments.

Test Difficulty

Only one study has investigated the effect of test difficulty on calibration. Pressley and Ghatala (1988) compared students' performance on opposites (selecting a word opposite in meaning from five alternatives), analogies (e.g., intruder is to privacy as ripple is to calm), and reading comprehension. The questions on each test were designed to range from easy to difficult based on pilot testing. Calibration was significant for all three measures. Results showed that students were more accurate at evaluating their performance when the test items were designated as easy rather than difficult.

A factor that may contribute to test difficulty is the type of information readers are required to recall on a comprehension test. It seems reasonable to assume that questions that require a reader to make inferences or to identify the theme of a passage are more difficult to answer correctly than questions that require mere recall of facts. Pressley, Ghatala, Woloshyn, and Pirie (1990) found that students were better at assessing their performance when questions on the comprehension test required recall of details rather than the main theme. The comprehension test consisted of practice texts from the Preliminary Scholastic Aptitude Test (*PSAT*) and the Scholastic Aptitude Test (*SAT*). Students then answered questions that related to particular details (e.g., the exact quantity of exports shipped to the United States from Japan) or required recognition of the overall theme (e.g., understanding the main point the author was trying to convey to the reader). Although performance on the comprehension test was not influenced by information type, students' postdictions were significantly better after answering detail questions compared to thematic questions.

Glenberg et al. (1987, Exp. 1), however, failed to obtain similar results when they compared students' calibration accuracy on inference questions or factual recall

questions. Students read 15 one-paragraph expository texts and, after reading each text, predicted how well they would do on a question over the material on a scale of 1 (low confidence) to 6 (high confidence). Similar to Pressley et al.'s (1990) findings, the type of information was not related to test performance. However, Glenberg et al. found that information type was also not related to calibration. Results showed that calibration was no better than chance for either inference (G = .14) or factual recall (G = -.13) questions.

Maki (1995, Exp. 1) reported mixed results when comparing students' performance on three types of questions. Maki asked students to predict their performance for what she termed important, unimportant, and higher order questions. Important and unimportant questions required recall of facts stated in the texts, whereas higher order questions required students to make inferences about information not directly found in the texts. Although Pressley et al. (1990) and Glenberg et al. (1987) reported that information type was not related to test performance, Maki found that students were significantly better at answering important and unimportant questions compared to higher order questions. Looking at prediction accuracy, Maki found, similar to Glenberg et al., that information type did not influence prediction accuracy in Experiment 1. However, in Experiment 2, Maki found, similar to Pressley et al., that information type was related to prediction accuracy. Students were better at predicting their performance when questions pertained to unimportant or higher order information.

Taken together, these findings suggest that calibration accuracy may be influenced by the type of information requested on the comprehension test. Students are sometimes, but not always, better at calibrating their performance on factual recall questions.

The issue of whether test difficulty influences adults' ability to accurately evaluate their comprehension performance is an important one. Equally important is the fact that test difficulty may confound results in studies investigating the effects of other variables. This point was made in a review of methodological issues in metacognition research when Schwartz and Metcalfe (1994) discussed what they referred to as an item selection effect. This occurs if one group of participants inadvertently ends up with a set of items that are easier or more difficult than another group. Any statistical differences observed between the two groups may then be due to differences in test difficulty rather than to differences in the calibration performance of the two groups.

Advantages of IRT

One of the major benefits of IRT is that, if an IRT model fits the test data, estimates of examinee ability and test difficulty are independent of one another (Hambleton, Swaminathan, & Rogers, 1991). This is not the case in CTT. In CTT, an individual's ability is defined as performance on a particular test. When the test is difficult, an examinee will appear to have low ability and, when the test is easy, will appear to have higher ability. Similarly, test difficulty is defined in terms of the scores obtained by the examinees taking the test. In CTT, a test will appear to be more difficult when examinees have low ability and less difficult when examinees have high ability. As a result, examinee ability and test difficulty can only be interpreted in relation to one another (Hambleton et al.).

In IRT, examinee ability and test difficulty are described by monotonically increasing functions called item characteristic curves (*ICCs*) (See Figure 1). Each test item is represented by an ICC. The probability of giving a correct response is plotted on

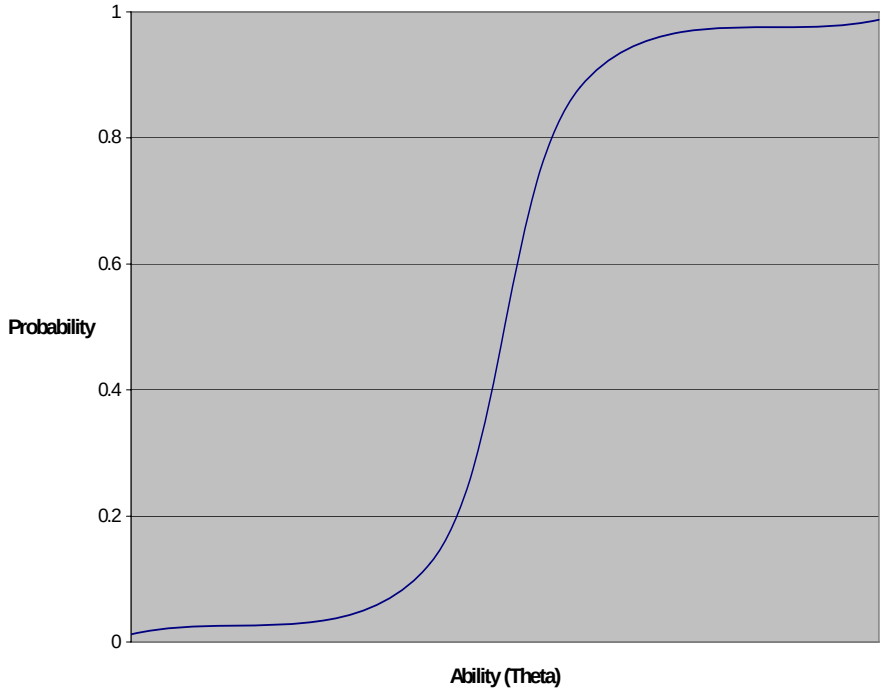

Figure 1. An item characteristic curve (ICC)

the vertical axis and theta (θ) is plotted on the horizontal axis (Rouse et al., 1999). Theta is a general term that refers to the underlying latent trait or ability that is being measured. The curve of an ICC describes how changes in ability level relate to changes in the probability of a correct response, and is determined by one or more item and ability parameters. Unlike CTT, there are a number of IRT models. Because there are assumptions associated with each model, the models must be tested to determine if they fit the data. If a model is found to fit, estimates of test difficulty and examinee ability are independent of one another (Hambleton & Swaminathan, 1985; Hambleton et al., 1991).

As a result, IRT may provide valuable insight into the role of test difficulty on postdiction accuracy. In CTT, item difficulty and discrimination depend on the test being

normed on a representative sample from the target population (Embretson, 1996). This can pose a problem for researchers investigating metacomprehension because they generally do not have ready access to representative samples but must choose participants from intact groups. For example, Schraw, Potenza, and Nebelsick-Gullet (1993) classified test items as easy if 80% of the students in a pilot study answered the item correctly and classified items as difficult if 50% answered the item correctly. However, unless the students in the pilot study and the students in the study were very similar in terms of ability, it cannot be presumed that the difficulty levels of the test items were comparable for the two groups. An IRT-based test yields unbiased estimates of item properties even from nonrepresentative samples so that the difficulty level of the test is independent of the participants taking the test (Embretson, 1996).

A second advantage is that IRT allows a more comprehensive description of a test's usefulness (Rouse et al., 1999). CTT-based tests are effective at discriminating between upper and lower halves of examinees but are ineffective at discriminating between examinees at other trait levels (Crocker & Algina, 1986). With the use of IRT, a test can be developed that is efficient at one or more specific levels of a trait or across the entire trait continuum.

An important practical consideration is that IRT allows for shortened testing sessions. Compared to a CTT-based test, an IRT-based test provides more information with fewer test items. This is likely to be an advantage to most researchers because they can develop reliable instruments with many fewer than 40 - 60 items as has been the case in some studies.

Because IRT allows for finer discriminations among test scores, an IRT-based test may help resolve mixed findings that have been reported about the effects of reading

ability on calibration accuracy. In CTT, each test item is given equal weight regardless of the item's difficulty level (Rouse et al., 1999). So an examinee who makes a correct response to an easy item is given the same score as an examinee who responds correctly to a difficult item. In 2PL and 3PL IRT models, items are not equally weighted so an examinee who responds correctly to a difficult item will not receive the same score as an examinee who responds correctly to an easy item. This benefit of IRT may be particularly useful for researchers who are investigating the effects of reading ability on monitoring accuracy because IRT-based test scores can provide more precise information about participants' reading ability than is possible with CTT-based assessments.

To summarize, an IRT-developed test offers clear advantages over the CTT-based tests that have been used in metacomprehension studies. Test difficulty and examinee ability are not tied to a particular test or to a particular group of test takers. Thus, the scores from different groups can be readily compared. Furthermore, IRT-based tests can be developed to assess readers' ability across a wide range of ability levels or to target a particular ability group. In addition, because items are not equally weighted, an IRT-based test offers a more precise estimate of a person's ability.

Unidimensional IRT models

Unidimensional IRT models are used when test performance can be explained by a single underlying latent trait. For example, a unidimensional model would be appropriate if performance on a comprehension test could be attributed to one main trait or ability (reading comprehension ability). Three frequently used unidimensional models are the one-, two-, and three-parameter logistic models. The mathematical form of the three-parameter (*3PL*) model is:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{[1 + \exp\{-Da_i(\theta - b_i)\}]}$$

where:

θ (theta) represents the latent trait,

P(θ) represents the probability of a correct response,

D is a scaling factor equal to 1.702, and

a, *b*, and *c* are the item parameters.

The two-parameter logistic (*2PL*) model can be obtained from the 3PL model by setting *c* = 0. The one-parameter logistic (*1PL*) model is obtained by setting *c* = 0 and *a* = 1 and eliminating the scaling factor *D* (Hambleton et al., 1991).

1PL model. The 1PL model (commonly known as the Rasch model) consists of the item difficulty parameter (*b*). Item difficulty in CTT is a value that can range from .00 to 1.00 and is defined as the proportion of examinees who answer the item correctly (Crocker & Algina, 1986). In IRT, *b* values generally range from -2.0 to 2.0 and indicate the point on the theta continuum at which an examinee has a 50% probability of making a correct response (Embretson & Reise, 2000; Hambleton et al., 1991). For example, if an item's *b* value = 1, an examinee with an ability score of one would have a 50% chance of making a correct response. An examinee with an ability score greater than one would have a greater than 50% chance of making a correct response, and an examinee with an ability score below one would have less than a 50% chance of making a correct response. The more that a person's ability exceeds or is less than the difficulty of the item, the

greater the probability that the person will pass or fail the item, respectively (Embretson, 1996).

2PL model. The 2PL model adds a scaling factor (D) and an item discrimination parameter (*a*) and indicates the steepness of the ICC. The *a* parameter is analogous to the CTT item-test correlation, and is appropriate when test items have varying levels of discrimination power. Large *a* parameters represent items that make fine discriminations among ability levels, and small *a* parameters indicate that examinees with low ability are as likely to respond correctly as examinees with higher ability. The typical values for *a* range from 0.00 to 2.00, with higher values indicating steeper slopes (Embretson & Reise, 2000; Hambleton et al., 1991).

3PL model. The 3PL model adds a pseudo-chance parameter (*c*) that takes into account the possibility that a person with low ability will answer a question correctly (Hambleton et al., 1991). This is most likely to happen when an examinee guesses the correct response. In the 1PL and 2PL models, the value of the *c* parameter is set at .00. In the 3PL model, *c* can range from .00 to 1.00, and indicates the probability level at which the ICC levels off (Embretson & Reise, 2000; Hambleton et al.).

Purpose of Study

Because of the potential benefits, the main purpose of this study was to develop an IRT-based comprehension test for use in future research. A second purpose was to investigate the possible effects of students' guessing on postdiction accuracy.

No research has been conducted to determine whether guessing on the comprehension test may influence adults' postdiction judgments. It seems reasonable to assume that a person who guesses on a test item is likely to have less confidence in the correctness of that response. The person may then report having little confidence that the

response was correct, which would indicate a high level of calibration accuracy. If the guess was the correct response, however, the participant's postdiction score would actually be lower as a result. The chance that a person will guess the correct response to a test item depends on the number of available alternatives on the test. If the item is a true-false format, the odds of picking a correct answer with no knowledge are quite high (Schwartz & Metcalfe, 1994). Examining the possible effects of guessing on postdiction accuracy seems warranted because a large number of calibration studies have been conducted using comprehension tests that have consisted of true-false items (e.g., Glenberg & Epstein, 1985; Glenberg et al.,1987; Lin, Moore, & Zabrucky, 2001; Lin, Zabrucky, & Moore, 1997; Magliano et al., 1993; Rawson, Dunlosky, & McDonald, 2002; Weaver, 1990).

Method

IRT analyses require substantially larger sample sizes than CTT analyses. Because the comprehension test used in the present study was taken from one used in a previous study conducted by Moore, Zabrucky, and Cummings (2003), it was possible to combine the comprehension test results of 435 students in the Moore et al. study with the test results of the 571 students who participated in this study. Therefore, IRT analyses of the comprehension test are based on a total sample of 1,006 adults. Findings with respect to postdictions and guessing scores, however, are based on data collected only from the 571 students who participated in this study.

Participants

Comprehension test. Participants were 1,006 adults with undergraduate or graduate education (*M* = 15.5 years). There were 799 women and 205 men in the sample (*M* age = 25.62 years, range = 18 to 58 years). Participants were recruited from two large

universities in the Southeast and received extra course credit for participating in the study.

Postdictions and guessing responses. A total of 571 students with undergraduate or graduate education (*M* = 16 years) participated in this part of the study. There were 448 women and 122 men in the sample (*M* age = 27.92 years, range = 18 to 59 years). Participants were recruited from a large university in the Southeast and received extra course credit for participating in the study.

Materials

Comprehension test. The comprehension test used in the Moore et al. (2003) study consisted of eight expository texts adapted from Glenberg and Epstein (1985). Each text was accompanied by six four-option multiple choice items. Because the postdictions and guessing responses in this study required substantially more time to answer than the postdictions in the Moore et al. study, only four of the eight texts were used to lower the risk that students would become tired. Data from the Moore et al. sample were analyzed, and the four texts that showed the greatest variability in response patterns were selected for use. (Appendix A contains the four texts and corresponding test items.) The mean number of words per text was 203 (range = 165 - 252). Flesch-Kincaid readability levels were 10.6, 12.0, 10.6, and 12.0, respectively. Texts were presented in a booklet form with the questions appearing on separate pages from the texts.

Postdiction. After each question on the comprehension test in the present study, there were instructions to respond *yes* or *no* to the following: "Are you more than 50% certain that you answered Item #___ correctly?" This postdiction was different than those used in previous studies for two reasons. First, it was hoped that this format would

provide more accurate information than has been possible with other types of postdictions. In the majority of calibration studies, participants have made confidence judgments on Likert scales. Because of the ordinal nature of Likert scales, such confidence ratings cannot be directly compared to percentages of correct performance on a comprehension test. As a result, the majority of studies have measured relative, rather than absolute, calibration (Lin, Zabrucky, & Moore, 2002). Determining absolute calibration would require an absolute relationship between confidence ratings and test performance (Lin et al., Nelson, 1996). To assess absolute calibration, some researchers have asked participants to report the number of items in a set they believe they answered correctly (e.g., de Carvalho Filho & Yuzawa, 2001; Lin et al., Pierce & Smith, 2001). For example, participants might read a passage, answer three corresponding test items, and report how many of the three items they believe they answered correctly. The drawback with this type of postdiction is that it does not permit an accurate assessment of calibration. To illustrate, suppose a participant gives the correct response to the first two test items, an incorrect response to the third test item, and then reports having confidence that two responses were correct. From this information, it is not possible to determine whether the participant was accurate (believed the responses to Items 1 and 2 were correct) or inaccurate (believed the responses to Items 1 and 3 or Items 2 and 3 were correct). Therefore, a different method to assess postdiction accuracy is needed.

Changing the postdiction also made it possible to use a different statistic to assess calibration accuracy. Several measures have been used to assess the relationship between performance and confidence judgments, and one area of debate has been which measure is the most appropriate (Keleman et al., 2000; Maki, 1998; Schwartz & Metcalfe, 1994). Among these are discrimination scores (Lundeberg, Fox, & Puncochar, 1994), bias scores

(Schraw, Dunkle, Bendixen, & Roedel, 1995), and correlation coefficients (Goodman & Kruskal's Gamma (*G*), Pearson's *r*, point biserial). Of these, gamma is generally considered the most appropriate statistic for use in metacognitive research (Nelson, 1984, 1996). Gamma is a nonparametric measure of association employed with ordinal data. Unlike other correlation coefficients, it is not interpreted in terms of variance accounted for, but has a probabilistic interpretation (Sheskin, 2000). Data from all possible pairs (confidence ratings and performance scores) are examined, revealing the probability that a text with a higher confidence rating will also have a higher performance score. Gamma ranges from –1 (if higher confidence ratings are paired with lower performance) to +1 (if higher confidence ratings are paired with higher performance).

Although widely used, gamma has several drawbacks. One drawback is that it treats only the rank of the underlying distribution of data so that the magnitude of differences cannot be determined (Schwartz & Metcalfe, 1994). Second, the use of gamma requires variability on both measures. If a participant makes either constant metacognitive judgments or exhibits constant criterion performance, a gamma correlation cannot be computed and that participant will be dropped from the analyses (see Glenberg & Epstein, 1987; Glenberg et al., 1987; Maki & Berry, 1984; Rawson, Dunlosky, & Thiede, 2000; Weaver & Bryant, 1995, for examples). As a result, participants with the highest comprehension performance and highest metacomprehension scores would be eliminated from the study as would participants with the lowest comprehension performance and lowest comprehension scores. This introduces a potential bias in the data.

Guessing. After each postdiction question, students were asked to respond to the following: (1) I chose the answer without guessing, or (2) I eliminated Option ___ and

guessed from the three remaining options, or (3) I eliminated Option ___ and Option ___ and guessed from the remaining two options, or (4) I eliminated Option ___, Option ___, and Option ____, and chose the remaining option.

Procedure

Students were tested in groups during their regular class meetings. At the beginning of the session, students were told that the purpose of the study was to determine whether adults can accurately assess how well they perform on a comprehension test. The test booklets were passed out, and students were asked to complete the first page, which requested demographic information. They were then asked to read the instructions on the second page of the booklet, which contained a sample test question, a sample postdiction rating, and a sample guessing response together with an explanation of how to fill out each. Students were then told that the test consisted of four short passages on different topics, and that they should read the passages carefully and take as much time as needed. They were also told that they should not turn back to a passage once they had completed it and turned the page to the multiple-choice questions.

Results

Comprehension

Comprehension accuracy was determined by summing the number of correct responses on the comprehension test. Scores could range from 0 to 20, with 20 indicating 100% accuracy. The mean comprehension score was 16.3 (SD = 2.86, range = 4 - 20). CTT statistics are presented in Table 3.

Table 3

Classical Test Theory Item Statistics

Item	No. Correct	% Correct	Item*Test Correlation Pearson	Item*Test Correlation Biserial
A2	945	.94	.21	.42
A3	759	.75	.24	.33
A5	884	.88	.41	.67
B1	853	.85	.26	.40
B2	865	.86	.35	.54
B3	713	.71	.33	.44
B5	721	.71	.33	.44
B6	697	.69	.36	.48
C1	908	.90	.24	.42
C2	879	.87	.31	.49
C3	738	.73	.33	.44
C4	929	.92	.38	.70
C5	449	.44	.26	.33
C6	889	.88	.46	.75
D1	885	.88	.39	.63
D2	834	.83	.32	.48
D3	895	.89	.44	.72
D4	756	.75	.36	.49
D5	455	.45	.25	.32
D6	900	.89	.33	.55

Note: Percentage correct = CTT item difficulty; biserial correlation = CTT item discrimination

Postdictions

Postdiction accuracy was computed by comparing students' responses on the comprehension test with their corresponding confidence ratings. A response was scored as accurate if students (1) made a correct response on the comprehension test item and responded that they were more than 50% sure they had answered the item correctly or (2) made an incorrect response on the comprehension test item and responded that they were not more than 50% sure they had answered the item correctly. Conversely, students were given a zero score if they (1) made a correct response on the comprehension test item and responded that they were not more than 50% sure they had answered the item correctly or (2) made an incorrect response on the comprehension test item and responded that they were more than 50% sure they had answered the item correctly. Postdiction scores could range from 0 to 20, with 20 indicating 100% accuracy. The mean postdiction score was 16.27 (*SD* = 2.44, range = 7 - 20). (See Table 4.) It was originally planned that a comparison be made between the Moore et al. (2003) postdiction ratings and the postdiction ratings in the present study. After further consideration, it was determined that it would be inappropriate to compare the two sets of ratings because the ratings were on different scales of measurement. In the Moore et al. study, students made a postdiction for each set of three comprehension test items, whereas in the present study students made a postdiction for each comprehension test item.

Demographic Variables

To determine if differences in gender, age, or education had an effect on performance, differences between group means were analyzed. Students were divided into younger (18 - 21 years) and older (22 - 58 years) age groups based on a median split. Students were also divided according to educational level, with one group consisting of

Table 4

Summary of Postdiction Scores

Score	*n*	%
7	1	.2
8	2	.4
9	3	.5
10	2	.4
11	15	2.6
12	28	4.9
13	21	3.7
14	65	11.4
15	59	10.3
16	69	12.1
17	100	17.5
18	97	17.0
19	81	14.2
20	28	4.9

Note: No scores = 0 - 7.

those who had 16 years or fewer of education, and one group consisting of those who had 17 years or more of education. T-tests revealed that all three demographic variables had a significant effect on both comprehension performance and postdiction accuracy. However, large sample sizes can increase the power of a statistical test and cause differences of even a very small magnitude to reach significance (Huck & Cormier, 1996). Therefore, effect sizes were computed to determine if the differences in means were practically significant. Effect size is a method of determining the standardized difference between two means that is independent of sample size (Cohen, 1988). According to Cohen, d = .2 is considered a small effect size, and d = .5 is considered a medium effect size. As can be seen in Tables 5 and 6, almost all of the effect sizes were small. The only variable that approached a medium effect size on the comprehension test scores was educational level. Students with at least 17 years of education scored an average of 1.31 points higher on the comprehension test than students who had 16 years or fewer education.

Guessing

A guessing score was computed for each participant. This was done by assigning values to the responses for the guessing questions that accompanied the 20 items on the comprehension test. The values were as follows: "eliminated one item" = .25; "eliminated 2 items" = .50; "eliminated 3 items" = .75; "did not guess" = 1. The guessing score could range from 5 to 20, with a score of 20 indicating that a student reported not guessing on any of the items. As can be seen in Figure 2, the majority of students reported that they did not guess on the comprehension test. The mean guessing score was 17.95 (SD = 1.75, range = 9.75 - 20). There was a significant relation between guessing scores and comprehension scores (r = .344, $p < .01$) and between guessing

Table 5

Comprehension Test Results for Demographic Variables

	N	*M*	*SD*	*t*	*d*
Gender					
Males	205	16.38	3.26		
Females	799	15.79	3.11	2.382	.185
Age					
18 - 21 years	440	15.44	3.33		
22 - 58 years	558	16.31	2.93	-4.368	-.277
Education					
≤ 16 years	705	15.51	3.29		
≥ 17 years	298	16.82	2.59	-6.095	-.442

Note: d = Cohen's d

Table 6

Postdiction Results for Demographic Variables

	N	*M*	*SD*	*t*	*d*
Gender					
Males	122	16.73	2.40		
Females	448	16.15	2.43	2.344	.240
Age					
18 - 21 years	152	15.72	2.37		
22 - 58 years	417	16.48	2.42	-3.349	-.317
Education					
≤ 16 years	276	15.84	2.50		
≥ 17 years	292	16.65	2.31	-4.026	-.336

Note: d = Cohen's d

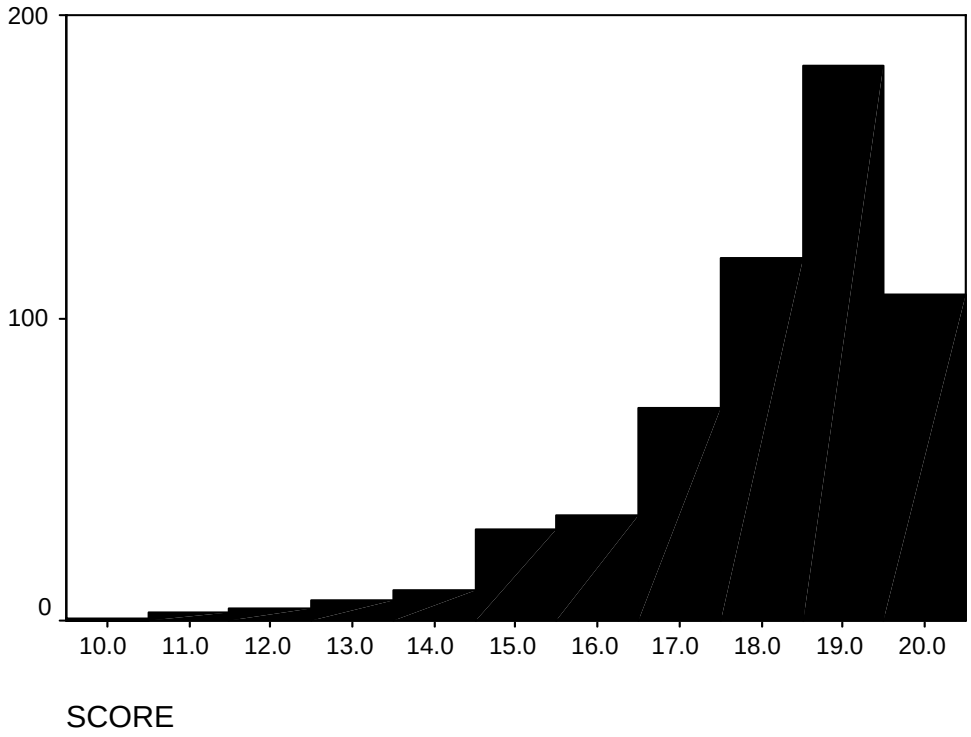

Figure 2. Histogram of students' guessing scores

scores and postdiction scores (r = .356, $p < .01$). Although the correlations were small, these findings suggest that students who were less likely to guess on comprehension questions demonstrated higher comprehension performance and higher postdiction accuracy.

IRT Results

A number of steps are required to assess the comprehension test using IRT. The steps are: (1) determine if model assumptions have been met, (2) identify which model is appropriate, (3) identify the ICC that best fits each test item, and (4) calculate ability estimates for the participants.

Model assumptions. Two assumptions that must be met when trying to fit unidimensional IRT models are unidimensionality and local independence (Hambleton et al., 1991; Lord & Novick, 1968). Although strict unidimensionality is unrealistic with any test instrument, the assumption of unidimensionality is generally considered met if a single dominant factor runs through the items (Embretson & Reise, 2000; Hambleton et al.; Lord & Novick; Stout, 1987). The assumption of local independence is considered satisfied if the assumption of unidimensionality is met (Hambleton et al.).

One widely accepted method for determining dimensionality is to examine the relative sizes of the eigenvalues associated with a factor analysis of the test items. To accomplish this, a matrix of tetrachoric correlations among the comprehension items was computed (Hambleton et al., 1991; Lord & Novick, 1968). A principal axis factor analysis with varimax rotation using SPSS (Version 11.5) resulted in eight factors being extracted. Evidence of a dominant single factor would be if the first factor accounted for approximately 20% or more of the variance and was several times larger than the second factor (Hambleton, 2004; Smith & Reise, 1998). In the present case, the first factor accounted for 27.08% of the variance, and there was a sharp decline in the eigenvalues from the first factor to the second and third factors (which accounted for 9.87% and 7.94% of the variance, respectively). These results are satisfactory for meeting the unidimensionality assumption.

A second widely used test of unidimensionality is the scree test (Bentler & Yuan, 1998; Reckase, 1979). To conduct a scree test, a plot is created with the number of factors on the x-axis and the corresponding eigenvalues (percentage of variance accounted for by a factor) on the y-axis. The scree plot shows the point where the eigenvalues form a downward trend (Reise, Waller, & Comrey, 2000). As can be seen in

Figure 3, a scree test resulted in a large eigenvalue for the first factor, which then dropped sharply to the second and third factors. This was taken as further evidence that performance on the comprehension test could be accounted for by a single dominant trait.

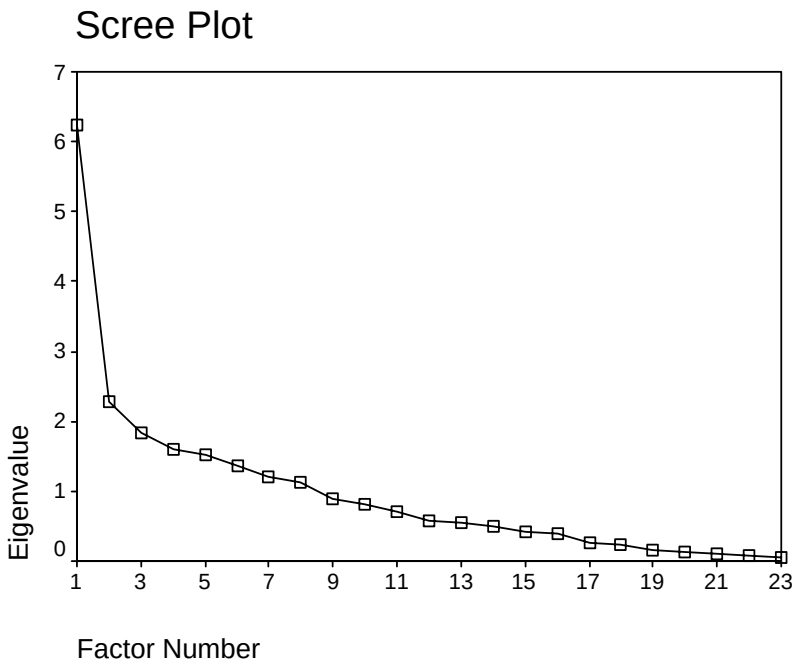

Figure 3. Results of scree test to determine unidimensionality

Model fit. All IRT analyses were performed using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). Default values were used with the exception that the program was specified to use up to 50 iteration cycles to come up with a solution.

Because the primary interest in this study was to fit an IRT model to an existing set of items, Embretson and Reise (2000) suggest that the 2PL or 3PL models are favored over the 1PL model. To determine model fit, the –2 log likelihood values (computed by

BILOG-MG) for each model can be compared. The differences in the values are evaluated using chi square statistics. The equation is:

$$X^2 = -2 \log \text{likelihood}_{2PL} - (-2 \log \text{likelihood}_{3PL})$$

The degrees of freedom are determined by the additional number of parameters of the more complex model. In the present study, -2 log likelihood for the 2PL model = 18390.2513, and -2 log likelihood for the 3PL model = 18369.0956. The change in X^2 from the 2PL to the 3PL model was 21.1557. The 3PL model added 20 degrees of freedom (for 20 pseudo-chance parameters). At 20 degrees of freedom, the difference was not significant. This indicated that the 3PL model did not provide a significantly better fit than the 2PL model. Therefore, all subsequent analyses were based on the 2PL model.

Item difficulty (b). Basically, easy items (those answered correctly even by persons with low trait levels) have negative *b* parameter values, and difficult items (items answered correctly only by persons with high trait levels) have large positive *b* parameter values. Results showed that item difficulty estimates ranged from –3.885 to .390. The majority of items were quite easy with only two having difficulty estimates greater than zero. Table 7 contains the item parameter estimates.

Item discrimination (a). The *a* values ranged from .325 to 1.005. Only one item (C6) had an estimate greater than 1.0. This finding indicates that none of the items on the comprehension test were very discriminating at any one point on the theta continuum.

Table 7

Final Item Parameter Estimates for the Comprehension Test

Item	*a*	*b*
A2	.474 (.070)	-3.885 (.486)
A3	.325 (.043)	-2.213 (.287)
A5	.815 (.094)	-2.037 (.145)
B1	.428 (.054)	-2.694 (.294)
B2	.605 (.075)	-2.224 (.195)
B3	.458 (.049)	-1.344 (.145)
B5	.484 (.051)	-1.346 (.135)
B6	.547 (.054)	-1.087 (.106)
C1	.458 (.066)	-3.279 (.393)
C2	.542 (.064)	-2.542 (.242)
C3	.502 (.054)	-1.430 (.138)
C4	.857 (.105)	-2.444 (.184)
C5	.379 (.044)	.390 (.113)
C6	1.005 (.107)	-1.893 (.111)
D1	.749 (.078)	-2.138 (.151)
D2	.538 (.060)	-2.103 (.193)
D3	.958 (.106)	-1.982 (.116)
D4	.560 (.057)	-1.451 (.133)
D5	.369 (.043)	.355 (.114)
D6	.651 (.075)	-2.478 (.198)

Note: Standard errors are in parentheses. a = item discrimination, b = item difficulty.

Item Analysis. BILOG-MG provides chi-square item-fit statistics which can be used to assess the fit of individual test items. These item-fit statistics are derived by sorting examinees into groups on the basis of their trait level and then comparing the observed proportion endorsed within a group with that predicted by the estimated item characteristic curve (ICC). One item (item A6) was dropped from the analysis because BILOG-MG would not converge when it was included. Examination of the CTT statistics revealed that the correlation between item A6 and the entire test was -.019 and the biserial correlation was -.024, indicating that the item was a poor one. With the criterion for item fit set at $X^2 < .0021$, three items (A1, A4, and B4) were judged not to fit with respective probabilities of .0001, .0004, and .0001. All analyses were conducted on the 20 remaining items.

Ability estimates. As can be seen in Table 8, students' ability levels ranged from –3.1712 to 1.5739. Because items did not have the same difficulty levels, students could have the same total CTT score, and receive different ability scores. For example, ability levels for students who answered 19 items correctly ranged from .9201 to 1.2034.

Discussion

The major goal of the present study was to develop a comprehension test using IRT. The results showed that a single underlying trait (reading comprehension ability) could account for performance on the comprehension test. A 2PL model was found to fit the data, suggesting that low-ability students did not make correct responses because they guessed on the items. If guessing had strongly influenced test performance, a 3PL model would likely have provided a better fit to the data than the 2PL model.

Table 8

IRT Ability Scale Scores

Comprehension Score	*n*	Ability Score	*M*	*SD*
4	1	-2.9085		
5	1	-2.7311		
6	2	-2.6460 to -2.6409	-2.644	.004
8	6	-2.3909 to -1.8221	-2.132	.215
9	8	-2.1654 to -1.6096	-1.837	.177
10	12	-1.7829 to -1.4972	-1.604	.087
11	13	-1.6010 to -1.3265	-1.443	.076
12	21	-1.5615 to -.9931	-1.298	.185
13	34	-1.4808 to -.6739	-1.062	.201
14	34	-1.1869 to -.4368	-.679	.169
15	55	-.7289 to -.1541	-.419	.142
16	62	-.5573 to .1283	-.130	.165
17	71	-.2482 to .5030	.248	.144
18	116	.2482 to .8350	.677	.124
19	90	.6005 to 1.2034	1.087	.109
20	45	1.5739	1.574	.000

Note: There were no comprehension scores = 1, 2, 3 or 7.

With respect to the test items, four of the 24 items were found to be ineffective at providing information about students' ability and were eliminated from the test. The results showed that the remaining 20 items were quite easy. Recall that *b* values generally range from -2.0 to 2.0. Only two of the 20 items had positive values (.390 and .355) and seven items had values that ranged from -1.09 to -1.98. Nine additional items had values that ranged from -2.0 to -2.6, and the remaining two items had values of -3.2 and -3.9. The test items also were not very discriminating. Values for the *a* parameter generally range from 0.00 to 2.00, with higher values indicating more discriminating items. The *a* estimates for the items ranged from .325 to 1.005 (M = .585) with only one item (C6) having an estimate greater than 1.0. Thus, the items were not very discriminating at any one point on the theta continuum, suggesting that lower-ability students were as likely to make correct responses as higher-ability students.

In sum, the comprehension test will be useful for distinguishing participants at very low ability levels from those at low ability levels, but will be ineffective at distinguishing among persons with average or high ability levels. In its present form, the comprehension test would be most appropriate to use when the target population is at the low end of the ability continuum (e.g., participants who are younger and less educated than adults in the present study).

Turning to the ability estimates, results showed that students' ability levels ranged from –3.1712 to 1.5739. Because IRT-developed test items do not have the same difficulty levels, two students could obtain the same total number correct, but receive different ability scores. So, for example, ability levels for students who answered 19 items correctly ranged from .9201 to 1.2034. It is important to point out that test scores are interpreted differently in CTT and IRT (Embretson & Reise, 2000). In CTT, it is

common practice to use norm referenced scores so that test scores can be interpreted only relative to the mean and standard deviation of a norm group. In IRT, test scores and item difficulties are on the same scale. This means that a person's ability score does not need to be interpreted relative to a norm group, but can be referenced directly to the test items. With respect to the comprehension test in this study, for example, it can be said that a **student with an ability estimate of 1.5739** is very likely to answer 100% of the items correctly because the highest item difficulty is .390. This difference in how scores are interpreted is likely to be of particular value to researchers who often lack the resources necessary to develop adequate norm referenced tests.

A second goal of this study was to examine the possible effects of guessing on postdiction accuracy. Although the vast majority of students reported that they did not guess on test items, this finding should not be taken to indicate that guessing is not a variable that warrants further investigation. It is possible that students did not have to guess because the comprehension test was so easy. Despite the low levels of guessing reported, there was a significant correlation between guessing scores and comprehension scores ($r = .344$, $p < .01$) and between guessing scores and postdiction scores ($r = .356$, $p < .01$). As one might predict, students who were less likely to guess on items obtained higher performance scores and were more aware of their performance. If there had been more variability with respect to difficulty on the comprehension test, the effects of guessing may have been more pronounced.

Although the main purpose of the present study was not to investigate the possible effects of gender, age, or education on comprehension performance and postdiction accuracy, effect sizes were calculated to determine any relationships. The results showed that none of the variables had an effect on postdiction accuracy, and the only variable that

had an effect on comprehension performance was educational level. This effect was of medium size, indicating that students who were in the process of earning a Master's degree scored higher than those who had not yet earned a Bachelor's degree. If the comprehension test had been more difficult, somewhat different findings may have been obtained.

It is noteworthy that students' postdiction accuracy was considerably higher than has been generally reported in the literature. Although postdictions are almost always significantly more accurate than predictions, they are nevertheless generally low (Pierce & Smith, 2001). However, the findings in the present study indicate that adults are quite good at judging how well they performed on a comprehension test. When students made a confidence rating for each of the 20 test items, their mean postdiction accuracy was 16.27 (81.4%). It is proposed that the design of the confidence judgment provided a more precise measure of postdiction accuracy than designs used in previous research.

Two limitations should be noted. A serious limitation is that there was not enough variability on the comprehension test to determine ability levels at the high end of the theta continuum. This made it difficult to make fine discriminations among students' ability and also made it difficult to determine the possible effects of guessing on students' postdiction accuracy. Future research should focus on the development of an IRT-based comprehension test that can accurately assess participants who are at the medium and high end of the ability continuum. This will make it possible to get a clearer picture of the effects of reading ability and test difficulty on adults' postdiction accuracy as well as provide further insight into the role that guessing may play.

A second limitation has to do with determining item fit. There is recent evidence to suggest that the use of the chi-square statistic to determine item fit is misleading (see

Neel, 2004). Because alternative measures were not available for use in the present study, item fit was determined using the chi-square statistic. Future research should explore alternative methods for assessing item fit. Additionally, researchers may want to examine models appropriate for polytomous data. The 1PL, 2PL, and 3PL models investigated in the present study are appropriate for use with dichotomously scored data. When items have more than two response options (as is the case with multiple choice tests), it is possible to dichotomize the responses by scoring one option as correct and all remaining options as incorrect (Chernyshenko et al., 2001). Because models appropriate for polytomous data have not been as extensively researched as those for dichotomous data, a decision was made to dichotomize the responses to the multiple-choice items. It may be advantageous to test polytomous models such as Bock's nominal response model, Samejima's graded response model, and Thissen and Steinberg's multiple-choice model (van der Linden & Hambleton, 1997).

Metacomprehension continues to be an important area of research. It has already been noted that there may be deficiencies in the testing procedures that have been widely used to investigate metacomprehension. The use of IRT techniques, together with a more precise measure of postdiction accuracy, are steps towards eliminating some of those deficiencies.

References

Baker, L. (1985). Differences in the standards used by college students to evaluate their comprehension of expository prose. *Reading Research Quarterly, 20,* 297-313.

Bentler, P. M., & Yuan, K. (1998). Tests for linear trend in the smallest eigenvalues of the correlation matrix. *Psychometrika, 63,* 131-144.

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36,* 523-562.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Commander, N. E., & Stanwyck, D. J. (1997). Illusion of knowing in adult readers: Effects of reading skill and passage length. *Contemporary Educational Psychology, 22,* 39-52.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Harcourt Brace Jovanovich.

De Carvalho Filho, M. K., & Yuzawa, M. (2001). The effects of social cues on confidence judgments mediated by knowledge and regulation of cognition. *Journal of Experimental Education, 69,* 325-345.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8,* 341-349.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 702-718.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15,* 84-93.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116,* 119-136.

Hambleton, R. K. (2004, June). *Traditional and modern approaches to outcomes measurement.* Paper presented at The National Cancer Institute (NCI) and Drug Information Association (DIA) conference, Bethesda: MD.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Huck, S. W., & Cormier, W. H. (1996). *Reading statistics and research* (2nd ed.). New York: HarperCollins.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258-272.

Keleman, W. L., Frost, P. J., & Weaver, C. A. III (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition, 28,* 92-107.

Lin, L., Moore, D., & Zabrucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology, 22,* 111-128.

Lin, L., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23,* 345-391.

Lin, L., Zabrucky, K. M., & Moore, D. (1997). The relations among interest, self-assessed comprehension, and comprehension performance in young adults. *Reading Research and Instruction, 36,* 127-139.

Lin, L., Zabrucky, K. M., & Moore, D. (2002). Effects of text difficulty and adults' age on relative calibration of comprehension. *American Journal of Psychology, 115,* 187-198.

Lord, F., & Novick, M. (1968). *Statistical theories of mental tests.* New York: Addison-Wesley.

Lundeberg, M. A., Fox, P. W., & Puncochar, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology, 86,* 114-121.

Magliano, J. P., Little, L. D., & Graesser, A. C. (1993). The impact of comprehension instruction on the calibration of comprehension. *Reading Research and Instruction, 32,* 49-63.

Maki, R. H. (1995). Accuracy of metacomprehension judgments for questions of varying importance level. *American Journal of Psychology, 108,* 327-344.

Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117-144). Mahwah, NJ: Erlbaum.

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., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 609-616.

Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin and Review, 1,* 126-129.

Maki, R. H., & Serra, M. (1992a). The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 116-126.

Moore, D., Lin-Agler, L. M., & Zabrucky, K. M. (2005). A source of metacomprehension inaccuracy. *Reading Psychology, 26,* 251-265.

Moore, D., Zabrucky, K. M., & Cummings, A. M. (2003). [Comprehension test data]. Unpublished raw data.

Neel, J. H. (2004). A new goodness-of-fit test for Item Response Theory. *Journal of Modern Applied Statistical Methods, 3,* 581-593.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109-133.

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. *Applied Cognitive Psychology, 10,* 257-260.

Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition, 29,* 62-67.

Pressley, M., & Ghatala, E. S. (1988). Delusions about performance on multiple-choice comprehension tests. *Reading Research Quarterly, 23,* 454-464.

Pressley, M., Ghatala, E. S., Woloshyn, V., & Pirie, J. (1990). Sometimes adults miss the main ideas and do not realize it: Confidence in responses to short-answer and multiple-choice comprehension questions. *Reading Research Quarterly, 25,* 233-249.

Rawson, K. A., Dunlosky, J., & McDonald, S. L. (2002). Influences of metamemory on performance predictions for text. *The Quarterly Journal of Experimental Psychology, 55A,* 505-524.

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28,* 1004-1010.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Reise, S. P., Smith, L., & Furr, R. M. (2001). Invariance on the NEO PI-R Neuroticism scale. *Multivariate Behavioral Research, 36,* 83-110.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12,* 287-297.

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 72,* 282-307.

Schraw, G., Dunkle, M. E., Bendixen, L. D., & Roedel, T. D. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology, 87*, 433-444.

Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology, 18*, 455-463.

Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J. A. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 93-113). Cambridge, MA: MIT Press.

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures* (2nd ed.). New York: Chapman & Hall.

Silver, B. B., Smith, E. V., Jr., & Greene, B. A. (2001). A study strategies self-efficacy instrument for use with community college students. *Educational and Psychological Measurement, 61*, 849-865.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *Journal of Personality and Social Psychology, 75*, 1350-1362.

Stark, S., Chernyshenko, O. S., Lancaster, A. R., Drasgow, F., & Fitzgerald, L. F. (2002). Toward standardized measurement of sexual harassment: Shortening the SEQ-DoD using item response theory. *Military Psychology, 14*, 49-72.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

van der Linden, W., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

Weaver, C. A. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 214-222.

Weaver, C. A., & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition*, *23*, 12-22.

Zabrucky, K. (1990). Evaluation of understanding in college students: Effects of text structure and reading proficiency. *Reading Research and Instruction, 29*, 46-54.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago, Scientific Software International.

APPENDIXES

APPENDIX A

Passage A. Viruses and Cancer

Today viruses are known to cause cancer in animals and in certain plants. Despite the fact, most people do not accept viruses as being of importance in human cancer. As we know, viruses are infectious agents. Since there is no evidence that human cancer is infectious, many people thus believe that viruses cannot possibly be significant causal agents in the development of human cancer. However, viruses can mutate.

Examples are known in which a virus that never kills its host can mutate to form a new strain of virus that always kills its host. It does not seem unreasonable to assume that a harmless virus might mutate to form a strain that causes cancer. Add to this argument the general observation that biological phenomena generally do not differ much as one goes from one species to another. If viruses have been found in the development of cancer in lower species, it is plausible to consider them as a cause of human cancer.

A1 Viruses have yet to be shown to cause cancer in _____.
- A. animals
- B. plants
- C. humans*
- D. all of the above

A2 Since there is no evidence that human cancer is infectious,
- A. people are afraid to shake hands with cancer patients
- B. people tend to think that viruses cannot be a causal agent in the development of cancer*
- C. people are quite indifferent about it
- D. none of the above

A3 Which of the following statements concerning viruses is not true?
A. viruses are known to cause cancer in animals
B. viruses are known to cause cancer in plants
C. viruses are infectious agents
D. viruses cannot mutate*

A4 Biological phenomena generally
A. do not differ much as one goes from one species to another*
B. varies widely from species to species
C. causes cancer
D. varies widely between humans and animals but not animals and plants

A5 If viruses have been found in the development of cancer in the lower species, it is reasonable to assume that
A. animals are more susceptible to cancer
B. animals are more susceptible to viruses
C. viruses could be a cause of human cancer*
D. monkeys cause cancer

A6 If a drug is developed to reduce the genetic mutation of viruses,
A. some forms of cancer may be prevented
B. viruses will not cause cancer in animals
C. viruses will not cause cancer in plants
D. viruses will not mutate to kill their hosts*

Passage B. Neutrinos and the Expanding Universe

Among the myriad subatomic particles, the neutrinos occupy a particularly intriguing position in both theoretical and experimental works. Theorists and experimenters are fascinated by the role of neutrinos and its effect on the expanding universe. Specifically, they are quite interested in determining the mass of neutrinos and whether or not neutrinos contribute to the gravitational process. As their name suggests, neutrinos are electrically neutral and they do not interact with charged particles. Further, they travel practically at the speed of light and have little, if any, mass; the result is that these particles interact very feebly with matter.

Thus, even though neutrinos are produced in copious amounts in the nuclear reactions which occur in the sun and other stars, it is quite difficult to detect their presence because they can pass easily through the Earth without losing a bit of energy.

However, if the neutrinos are not completely massless, then by virtue of their immense population they should contribute significant gravitational influence on all the matter in the universe. The universe is known, on the large scale, to be expanding, with each region of matter moving away from all others. However, the gravitational attraction, by virtue of the mass of the neutrinos in the universe, may have sufficient pull to halt and even reverse this expansion. Because the role possessed by neutrinos may very well produce marked consequences in the universe, tremendous effort has been put forward by enthusiastic researchers to investigate the gravitational effect of neutrinos.

B1 Which of the following statements about neutrinos is(are) true?
A. neutrinos are subatomic particles
B. neutrinos move at the speed of light
C. neutrinos have little, if any, mass
D. all of the above are true*

B2 The charge of a neutrino is
A. positive
B. negative
C. neutral*
D. cannot be determined

B3 When passing through matter, neutrinos
A. react strongly with matter
B. react feebly with matter*
C. do not react at all with matter
D. interaction with matter cannot be determined

B4 Neutrinos are produced by
A. the nuclear reactions that occur in the sun and stars*
B. intense gravitation pull
C. the expansion of the universe
D. small, beaver-like animals

B5 When passing through the Earth, neutrinos
A. become positively charged
B. become negatively charged
C. do not lose any energy*
D. none of the above

B6 The mass of the neutrinos
A. exert no gravitational pull
B. may contribute an immense gravitational pull*
C. is causing the universe to expand
D. produce minimal consequences in the universe

Passage C. A Good Hanging Never Hurt Anyone

Capital punishment is popular once more. In fact, the news reports convey the sense that there is a driving impatience to get on with it. In the current state it will be well to remind ourselves of one important reason to stall the rush toward capital punishment. An examination of the judicial process by which society chooses who is to die shows that the process is imperfect. Instances of mistake are commonplace. These imperfections may be inevitable in any process concerning the emotion-laden charge of murder.

A decision to begin an act which can neither be reversed nor offset by compensating actions, on the basis of results of an imperfect process, is highly questionable. Inasmuch as death is irreversible and cannot be stopped by other actions, capital punishment should be avoided. It is, of course, true that in some sense even imprisonment is irrevocable. A day without freedom cannot be given back in kind. But some actions can be done, and the law makes remedial actions in the case of error. Advocates of capital punishment should recognize that carrying out an irrevocable act for which there is no possible compensation requires a degree of certainty and perfection of process that cannot be found in the judicial system.

C1 According to the text,
A. most people are against capital punishment
B. capital punishment has become popular once more*
C. people who are charged with murder should be sentenced to death
D. none of the above

C2 Examination of the judicial process indicates that in capital punishment cases, mistakes
A. are rare
B. are irrelevant to the judicial process
C. are commonplace*
D. are never serious enough to cause real problems

C3 Decreasing the likelihood of judicial error in capital punishment cases would result in
A. stronger arguments for capital punishment*
B. the abolishment of capital punishment
C. overall indifference to capital punishment
D. an increase in capital crime

C4 Which of the following does the paragraph describe as an argument against capital punishment?
A. it is not popular
B. mistakes are not commonplace
C. capital punishment is irreversible*
D. if a mistake occurs, it is easy to remedy

C5 An argument presented in the text indicates that people who have been wrongly jailed
A. can be compensated in some way*
B. are compensated sufficiently with freedom after the error has been discovered
C. can never be compensated
D. should probably be kept in jail for a while to make certain an error was made

C6 Based on the text, which of the following statements best describes the author's point of view on capital punishment?
A. Capital punishment should be advocated.
B. The practice of capital punishment has numerous benefits.
C. Capital punishment should be used on repeated offenders only.
D. Capital punishment should be avoided.*

Passage D. Rising Sea Levels

Scientific investigators of global climate change have cautioned that there will occur substantial rises in world-wide sea levels if there is a rise of several degrees in global temperature. The anticipated rise of global temperatures is attributed to the fact that since the middle of the 19th century personal and industrial uses of carbon-dioxide-releasing, combustible fuels have been increasing. The carbon dioxide is delivered and penetrates into the earth's atmosphere where it acts somewhat like the glass in a

greenhouse and absorbs infrared heat radiation from the earth instead of allowing it to escape into space.

Trapping of the infrared radiation will result in rising temperature. Even an increase of a few degrees of global temperature may produce melting of the polar ice caps and considerable inflows to the seas. The ensuing rises in sea level could have catastrophic consequences for the heavily inhabited coastal areas around the globe. Since it is not reasonable to relocate large populations away from the coastal cities, there is need to consider other actions in order to mitigate the danger to the future.

D1 According to the text, which of the following factors has contributed to the rise of global temperature?
A. forest fires
B. water shortage
C. the use of combustible fuels*
D. all of the above

D2 Carbon dioxide affects global atmosphere by
A. creating more oxygen
B. releasing heat
C. cooling the air
D. absorbing heat but not allowing it to escape*

D3 All of the following are problems associated with global warming except
A. melting of polar ice caps
B. declining of sea levels*
C. rising of sea levels
D. rising of temperatures worldwide

D4 The effect of global temperature changes of only a few degrees can be described as
A. modest
B. unpredictable
C. somewhat serious
D. catastrophic*

D5 According to the text, relocating coastal populations would be
A. advisable
B. necessary
C. impossible
D. unreasonable*

D6 Based on the text, the rise in sea level can be produced by
A. rain
B. earthquake
C. melting of polar ice caps*
D. none of the above