Effects of text difficulty and adults' age on relative calibration of comprehension

LIN-MIAO LIN
University of Southern Mississippi-Gulf Coast

KAREN M. ZABRUCKY
Georgia State University

DEWAYNE MOORE
Clemson University

The American Journal of Psychology
Summer 2002, Vol. 115, No. 2, pp. 187–198

Effects of text difficulty and adults' age on relative calibration of comprehension

LIN-MIAO LIN
University of Southern Mississippi-Gulf Coast

KAREN M. ZABRUCKY
Georgia State University

DEWAYNE MOORE
Clemson University

Weaver and Bryant (1995) proposed the optimum effort hypothesis, suggesting that undergraduate students were better able to predict comprehension when text materials matched their reading level (grade 12) as opposed to being too easy or too difficult (Weaver and Bryant did not assess the actual reading level of their participants). In the study, we examined the reading level and accuracy of performance prediction of both younger and older adults using Weaver and Bryant's materials. Regardless of our participants' high reading levels (grade 14 and above), they still predicted performance best when texts were written at around the grade 12 level, failing to support the optimum effort hypothesis.

According to Baker and Brown (1984), metacognition consists of two components: knowledge about cognition and regulation of cognition. An individual's knowledge about cognition is his or her accumulated knowledge about general cognitive processes, tasks, and strategies. Regulation of cognition, on the other hand, is the ongoing tracking of mental processing and use of regulatory strategies to facilitate cognitive performance. These two metacognitive components are "somewhat separate phenomena," (Baker & Brown, 1984, p. 353) with the former being more stable than the latter. For example, a person who has knowledge about comprehension processes and effective comprehension strategies is likely to maintain the information and yet may not actually apply the knowledge to evaluate comprehension during reading or use strategies when needed.

Researchers examining metacognitive skills in comprehension (i.e., metacomprehension) investigate both the knowledge and regulation components. Knowledge about comprehension often is assessed by questionnaires that measure an individual's self-reported knowledge

AMERICAN JOURNAL OF PSYCHOLOGY
Summer 2002, Vol. 115, No. 2, pp. 187–198

about comprehension processes and effective reading techniques (Moore, Zabrucky, & Commander, 1997a; 1997b). The regulation component of comprehension examines the extent to which readers evaluate understanding of texts and use strategies to regulate comprehension (Zabrucky & Moore, 1994). One research paradigm widely used to study regulation of comprehension, called calibration of comprehension (Glenberg & Epstein, 1985), considers the accuracy with which readers evaluate the final state of their comprehension without specifically examining their use of regulatory strategies during reading. The procedure used in calibration of comprehension research is to have participants read a text and assess how well they think they understand the text or rate their confidence in their ability to answer questions about the text just read. Calibration of comprehension is the ability to accurately assess comprehension states and is determined by the relationship between confidence (or understanding level) and performance.

In this type of research, understanding and confidence judgments often are made on Likert scales. According to Nelson (1996), a predicted numerical value on a Likert scale does not actually correspond to a particular percentage of correct performance on the criterion test because of the ordinal nature of Likert scales. Thus, this type of measure indicates relative accuracy of performance prediction rather than absolute calibration because the latter refers to the absolute relationship between the predicted value and actual performance value on the test. Therefore, *calibration of comprehension* technically is a misnomer used by Glenberg and Epstein (1985) and later researchers. To describe this type of research more accurately, we will use the term *relative calibration* to capture the research focus on the relative accuracy of predicting performance and to distinguish this type of research from the study of absolute calibration.

The study of relative calibration of comprehension, initiated more than a decade ago by Glenberg and Epstein and their colleagues (Glenberg & Epstein, 1985; Glenberg, Wilkinson, & Epstein, 1982), has continued to attract research interest. Several participant, task, and text variables appear to influence readers' ability to predict comprehension (see Lin & Zabrucky, 1998, for a review). For example, individual characteristics such as readers' domain knowledge and general ability (Glenberg & Epstein, 1987; Glover, 1989), metacognitive knowledge (Bouffard-Bouchard, 1994), and motivational orientations (Kroll & Ford, 1992) can influence relative calibration accuracy, as can task variables such as test formats (Glenberg, Sanocki, Epstein, & Morris, 1987; Maki & Serra, 1992a), retention intervals (Glenberg & Epstein, 1985; Glenberg et al., 1987), and feedback and practice (Maki & Serra, 1992b;

Walczyk & Hall, 1989). Although few text variables have been examined, Weaver and colleagues have recently found that text difficulty is related to relative calibration (Weaver & Bryant, 1995; Weaver, Bryant, & Burns, 1995).

Results of early studies of relative calibration of comprehension (Glenberg & Epstein, 1985, 1987) suggested that readers were not able to predict comprehension, as shown by low correlations between confidence ratings (or understanding ratings) and test performance scores (gamma correlations were never higher than .20).[1] These low correlations often were a result of readers' overestimation of how much had been comprehended. This overconfidence was called illusion of knowing by Glenberg et al. (1982).

More recently, however, investigators have been able to show that readers can reliably predict their comprehension, although correlations between confidence and performance still tend to be low (e.g., mean gammas of less than .35). Researchers in early studies often used a single test item per text to assess comprehension (Glenberg & Epstein, 1985, 1987), a methodological flaw that clearly constrained readers' comprehension evaluation ability. By constructing multiple questions for each text, Maki, Foley, Kajer, Thompson, and Willert (1990) and Weaver (1990) were able to detect a significant increase in readers' relative calibration level, thus demonstrating that the measurement method plays a role in the level of relative calibration accuracy demonstrated by individuals.

As researchers continued to explore different variables that might constrain readers' ability to successfully evaluate comprehension, a higher than usual relative calibration level was recently reported by Weaver and Bryant (1995), who showed that correlations of relative calibration reached as high as .69 when some text variables were taken into consideration. Weaver and Bryant (1995, Experiment 2) examined the role of text difficulty in relative calibration of comprehension, using narrative and expository texts at three readability levels. These levels were "easy" (below grade 8), "standard" (around grade 12), and "difficult" (around grade 16). The majority of students in the study consisted of first- and second-year college students from introductory psychology classes. Although Weaver and Bryant (1995) did not examine reading ability directly, they assumed that the standard passages best matched the reading ability of their participants.

Results of their study revealed an inverted U-shaped pattern of relative calibration accuracy as a function of text difficulty. Readers of standard passages were able to predict comprehension with high accuracy (the gamma between understanding ratings and performance was .69). However, students who read either easy or difficult texts were much less

capable of predicting comprehension (G = .29 and .30, respectively). Weaver and Bryant concluded that because the students were more likely to read at a grade 12 level, they were able to evaluate comprehension with the greatest accuracy when text materials matched their functional reading level.

Weaver and Bryant thus proposed the optimum effort hypothesis as one explanation of their findings. They suggested that the greatest relative calibration ability should occur when one is reading texts at a level that necessitates an intermediate amount of cognitive effort. Fairly easy materials, on the other hand, may direct readers to adopt an "automatic reading mode" (1995, p. 19) because of minimum demands on cognitive resources, whereas the amount of cognitive effort needed to process comparatively difficult texts may exceed an individual's available capacity, leaving no resources for adequate metacomprehension processing. In sum, an optimum calibration level is not likely to occur when too easy or too difficult texts are encountered relative to an individual's reading ability.

In the past, researchers studying relative calibration of comprehension often have used the texts developed by Glenberg and Epstein (1985). Although most of these texts were written at beyond grade 16 reading levels, researchers generally examined calibration in beginning-level college students, who were likely to read at a level lower than grade 16. The discrepancy between the difficulty of texts and students' reading ability might explain the low relative calibration levels often detected in those early studies.

A primary purpose of the present study was to further examine Weaver and Bryant's (1995) optimum effort hypothesis. Instead of college students at the beginning level, only individuals with a graduate degree or at least some graduate education were recruited to participate in our study. Those participants were likely to have a functional reading level of grade 16 or above (their actual reading level was determined by a standardized reading test). In the study, we used the texts developed by Weaver and Bryant (1995, Experiment 2), which were written at three difficulty levels. The optimum effort hypothesis suggests that the greatest relative calibration accuracy for our participants should occur when reading Weaver and Bryant's (1995) difficult texts rather than standard texts because the texts presumably would match our participants' functional reading level in the same manner that the standard texts presumably matched the reading level of Weaver and Bryant's participants. In our study, unlike that of Weaver and Bryant (1995), we directly examined participants' reading ability by a standardized test.

Another purpose of the present study was to investigate whether adults' age influences their ability to evaluate comprehension. In near-

ly all investigations of relative calibration of comprehension, participants have been young adults with educational backgrounds at the beginning college level. In the present study, we not only included an examination of relative calibration in more educated adults but also assessed relative calibration in younger as well as older adults. Although the available evidence on age-related differences in adults' metacomprehension and metamemory ability suggests that older adults, in general, are as accurate as young adults in metacognitive evaluation (Hertzog, Dixon, & Hultsch, 1990; Olin & Zelinski, 1997; Zabrucky & Moore, 1994; Zabrucky, Moore, & Schultz, 1993), very little research has been done using the relative calibration paradigm. Thus, our study was designed to reveal whether younger and older adults would exhibit similar metacognitive skills using a relative calibration paradigm as they have for other previously investigated paradigms.

EXPERIMENT

METHOD

Participants

Participants were 60 younger (mean age = 26.63 years, range = 23 to 35; mean education = 17.73 years, range = 17 to 20) and 60 older (mean age = 70.32 years, range = 61 to 84; mean education = 18.25 years, range = 17 to 20) adults with graduate education or a graduate degree. Younger participants were graduate students recruited from four universities in the southeast. Older participants consisted of church members, university alumni, and students attending local senior universities. Participants were given an honorarium for participating.

No significant differences were found for self-reported current health and vision on 7-point scales (1 = *poor* and 7 = *excellent*) between younger (*M*s = 6.20 and 6.20, respectively) and older participants (*M*s = 6.09 and 5.99, respectively). However, the older group had lower self-reported hearing acuity, $t(117) = 2.07$, $p < .05$ ($M = 6.13$ for younger adults; $M = 5.71$ for older adults).

Materials

Relative calibration of comprehension task. Twelve texts developed by Weaver and Bryant (1995, Experiment 2) were used in the experiment. Weaver and Bryant used the Flesch (1951) readability scale to determine text difficulty and divided texts into three sets of four passages. The sets were easy (reading level below grade 8), standard (reading level around grade 12), and difficult (reading level around grade 16). In the present study, we used the same terms Weaver and Bryant (1995) used to label the three sets of texts (i.e., *easy, standard,* and *difficult*).

Each participant in our study read one of the three sets of passages; that is, they read a set of either four easy, four standard, or four difficult passages. In each set, there were two expository and two narrative passages.[2] In the easy set,

narrative passages were "Old Sultan" and "The Wolf and Seven Young Kids," and expository passages were "The Universe" and "Euripides." In the standard set, expository passages included "The Martian Atmosphere" and "Symbiosis," and narrative passages were "Clever Answers" and "The Owl." In the difficult set, narrative passages included "The Queen Bee" and "Charley," and expository passages consisted of "The Heritage of Pharmacology" and "The Stock Market."

Each passage was divided into four paragraphs, with each paragraph containing approximately the same amount of reading material. The same set of multiple-choice questions constructed by Weaver and Bryant (1995) was used in the present study. Sixteen questions were generated from each passage; four questions were devised from each paragraph. Each participant answered a total of 64 questions across four passages. At the end of each passage, a 6-point Likert scale was prepared for participants to judge how well they understood the passage.

Vocabulary task. We assessed younger and older adults' vocabulary ability with the vocabulary subscale of the Nelson–Denny Reading Test Form E (Brown, Bennett, & Hanna, 1981). The vocabulary test contained 100 multiple-choice items. Participants were asked to choose from five options the word that best described the meaning of an underlined word in each opening statement. They were given 15 minutes to work on the test.

Reading comprehension task. We assessed younger and older adults' reading comprehension ability with the comprehension subscale of the Nelson–Denny Reading Test Form E (Brown et al., 1981). There were eight passages in the comprehension test, with each passage followed by several multiple-choice questions (a total of 36 questions was generated). Participants were asked to read completely through a passage and then answer the questions concerning the passage just read. Participants were allowed 20 minutes to work on the test.

Procedures

Participants were tested individually or in small groups of no more than four people in each group. The experimental session consisted of two parts. In the first half of the session, participants were asked to provide some demographic information and rate their health, vision, and hearing on Likert scales. They were then given the comprehension and vocabulary subscales of the Nelson–Denny Reading Test Form E (Brown et al., 1981) along with two other timed tasks not discussed in the present article. The order of tasks was randomized.

The relative calibration of comprehension task was administered in the second half of the session (followed by another task not discussed in the present article). Participants in each age group were randomly assigned to one of three conditions (each consisting of 20 individuals) to read four easy, four standard, or four difficult texts. Participants in all three conditions had equivalent reading ability, as measured by the Nelson–Denny comprehension test. The raw score means were $M = 27.65$, $SE = .79$ for readers of easy texts; $M = 27.39$, $SE = .81$ for readers of standard texts; and $M = 28.25$, $SE = .79$ for readers of difficult texts; maximum score = 36. Participants' verbal ability, as measured by the Nelson–Denny vocabulary test, was equivalent across text conditions and age

groups. The raw score means were $M = 91.88$, $SE = 1.29$ for readers of easy texts; $M = 91.73$, $SE = 1.29$ for readers of standard texts; and $M = 92.70$, $SE = 1.29$ for readers of difficult texts; $M = 91.70$, $SE = 1.05$ for younger adults; $M = 92.50$, $SE = 1.05$ for older adults; maximum score = 100. Younger participants had significantly better performance on the Nelson–Denny comprehension test, $M = 31.43$, $SE = .64$, than older participants, $M = 24.09$, $SE = .66$. Raw scores on the Nelson–Denny comprehension test were converted into grade-equivalent scores. Younger adults had an average reading level of grade 16.4, and older adults had a reading level of grade 13.7. Thus, the younger adults read at a grade level that was closer to that of the difficult texts, whereas the older adults read at a grade level that fell in between the standard and difficult texts.

On the relative calibration of comprehension task, participants were told that they were to work through the passages at their own pace. They were encouraged to read carefully and to take as much time as needed. For consistency with Weaver and Bryant's (1995) study, participants were asked to read each passage once for comprehension and were not informed about the comprehension test. Immediately after reading each text, participants were asked to rate how well they understood a passage on a 6-point Likert-type scale (1 = *very poorly,* 6 = *very well*) and also provide two other ratings not discussed in the present article. Presentation order of ratings was counterbalanced. Participants were not allowed to make changes on the ratings once they started reading a new passage.

After participants read all four passages, they were asked to answer 64 multiple-choice questions derived from the just-read texts. For consistency with the Weaver and Bryant (1995) study, participants were presented with 16 questions on a specific passage as a set, with each set being titled (e.g., "You will now be asked some questions over the passage titled . . ."). The order of the passages read and the sets of questions answered were randomized, as were the questions within each set. Participants were not allowed to reexamine texts during the comprehension test.

RESULTS

To examine relative calibration of comprehension on the Weaver–Bryant passages and to be consistent with the Weaver–Bryant procedure, we summed the four comprehension questions for each paragraph for each passage so that each passage had four comprehension scores (one for each paragraph). These four scores were then paired with the understanding ratings for each passage. Thus, each participant had four pairs of comprehension scores and ratings for each passage, resulting in a total of 16 data points across four passages. Gamma correlation coefficients were computed for each participant, across all texts, to determine the relationship between understanding ratings and comprehension performance on the multiple-choice questions. The mean gamma correlation coefficients across participants were calculated as a function of text difficulty.

A 2 (age) × 3 (text difficulty level) analysis of variance was conducted to investigate the effects of age and text difficulty on relative calibration accuracy. The results failed to show a main effect of age, $F(1, 105) = .30$, $p > .05$, $MSE = .17$, or a significant interaction between adults' age and text difficulty, $F(2, 105) = .29$, $p > .05$, $MSE = .17$. Thus, older adults were as able as younger adults to predict comprehension and were affected by text difficulty in a way similar to younger adults.

The main effect of text difficulty was significant, $F(2, 105) = 17.11$, $p < .0001$, $MSE = .17$. Fisher's least significant difference test revealed that the relative calibration means were significantly different between text conditions. The relative calibration mean was higher for the standard condition, $G = .47$, $SE = .07$, $N = 40$, than either the easy, $G = .19$, $SE = .07$, $N = 34$, or difficult, $G = -.09$, $SE = .07$, $N = 37$, conditions, with the latter two gamma coefficients being different as well. The results thus showed that readers' relative calibration accuracy varied as a function of text difficulty. Like Weaver and Bryant (1995), we found that readers evaluated comprehension more accurately with the standard texts than either the easy or difficult texts. However, we failed to find support for Weaver and Bryant's optimum effort hypothesis. Younger readers of difficult texts, whose functional reading level was most similar to the reading level of the difficult texts, still evaluated comprehension more accurately when reading the standard texts. Indeed, both younger and older adults, although having functional reading levels that were higher than the reading level of the standard texts, evaluated comprehension more accurately when reading the standard texts than when reading the difficult or the easy texts.

DISCUSSION

One major goal of the present study was to examine the effect of text difficulty on readers' ability to evaluate comprehension. Weaver and Bryant (1995) proposed the optimum effort hypothesis, which suggests that readers' ability to predict comprehension depends on their reading level relative to the difficulty of texts. Specifically, readers should be able to predict their comprehension most accurately when the difficulty level of the text matches their functional reading level. Results of the study revealed an inverted U-shaped pattern of relative calibration accuracy as a function of text difficulty. Accuracy of relative calibration was higher for standard passages than for easy or difficult passages.

The average reading level of our younger participants was grade 16. Weaver and Bryant's (1995) difficult passages, which were written at about the grade 16 level, most closely matched the reading level of the younger participants. Thus, according to the optimum effort hypothesis, young-

er readers should have been best able to predict comprehension of the difficult texts. However, they actually had poorer relative calibration for the difficult passages than for either easy or standard passages, which were written well below their reading level. Older participants, who had an average reading level of grade 14, showed a pattern of relative calibration of comprehension similar to that of younger participants. They also predicted comprehension best on standard passages and worst on difficult passages. However, in their case the standard passages were below their reading level, whereas the difficult passages were above their reading level. Thus, in the present study, a match between reading ability and text difficulty level did not guarantee optimal relative calibration accuracy. Participants evaluated comprehension best on Weaver and Bryant's standard passages regardless of their reading level. Thus, Weaver and Bryant's (1995) optimum effort hypothesis was not supported.

It may be the case that even very proficient readers might have trouble evaluating their comprehension of difficult passages. As noted by Weaver and Bryant (1995), the measure of readability used in their study to assign difficulty levels was simplistic, and the definition of "difficult texts" was not well established. Texts with high difficulty levels often include the use of unfamiliar words, long, complicated sentences, novel content, ambiguous causal relationships, and the use of abundant inferences that may entail knowledge beyond that given in the text. Such text characteristics may hinder even a skilled reader's ability to evaluate comprehension. For example, readers may not be able to judge comprehension accurately when they encounter a text with unclear cause–effect relationships that rely on extensive knowledge of a topic that they do not possess. Given our failure to support Weaver and Bryant's (1995) optimum effort hypothesis, it seems important to determine whether individuals of different reading levels continue to find an advantage for texts written at Weaver and Bryant's "standard" level (i.e., around grade 12 level), using completely different materials than those used by Weaver and Bryant.

Although our participants had problems judging their comprehension of the difficult texts, they comprehended all of the texts equally well. Participants' comprehension scores were equivalent across text conditions (63%, 60%, and 60% accuracy rates for easy, standard, and difficult texts, respectively). This finding differed from that of Weaver and Bryant (1995), in which comprehension scores were higher for easy passages than for either standard or difficult passages (61%, 49%, and 46% accuracy rates on easy, standard, and difficult texts, respectively). Thus, compared with the less-skilled readers in Weaver and Bryant's (1995) study, our skilled readers had better comprehension for standard and difficult texts.

In the present study, we also examined age-related differences in relative calibration of comprehension. Studies of relative calibration have focused almost exclusively on skills in younger adults, and ours is one of the first studies to examine relative calibration skills in younger and older adults. Results of our study indicate that older adults (at least those who are highly educated) are as able as younger adults to evaluate comprehension. The presence of similar relative calibration accuracy for younger and older adults found here is consistent with earlier studies on age-related differences in adults' ability to evaluate their understanding during reading. For example, using an error detection paradigm we found that older adults often are as able as younger adults to evaluate their understanding of the texts they are reading (Zabrucky & Moore, 1994) as indicated by the detection of specific problems embedded in texts. However, we also have found that older adults show problems in some metacognitive activities during reading, most notably in regulating their understanding with strategies such as rereading. Thus, older adults are as able to detect problems but do not selectively reread the problems to the same extent as younger adults. The results of the present study suggest that older adults' relative calibration skills are quite similar to those of younger adults and that their skills are influenced by text difficulty in similar ways as well.

Notes

Portions of these data were presented at the annual meeting of the Southeastern Psychological Association, Mobile, Alabama, 1998, and the annual meeting of the American Psychological Society, Washington, D.C., 1998. This research was supported by a grant from the National Institute on Aging (R29 AG09208) to Karen M. Zabrucky. We would like to thank Charles A. Weaver III and Deborah S. Bryant for supplying us with the passages and questions used in this research.

Correspondence about this article should be addressed to Lin-Miao Lin, Division of Education and Psychology, University of Southern Mississippi–Gulf Coast, Gulfport, MS 39560 (e-mail: lin-miao.lin@usm.edu). Received for publication April 26, 2000; revision received October 20, 2000.

1. Relative calibration accuracy generally is calculated by computing the gamma correlation between confidence ratings and comprehension performance. Nelson (1984) has shown that the gamma correlation coefficient is the appropriate correlational statistic for comparing the accuracy of predicting performance between two items. Data from all possible text pairs are examined when a gamma coefficient is computed, revealing the probability that a text with a higher comprehension rating will also have a higher performance score. The gamma statistic ranges from −1 (if higher ratings are always paired with lower performance) to +1 (if higher ratings are always paired with higher performance).

2. Although two types of texts (narrative and expository) were used, because of an inadequate number of observations, we were not able to test the effect of text genre on calibration of comprehension. Significantly more texts would be needed in each text difficulty condition to examine the effect of genre.

References

Baker, L., & Brown, A. (1984). Metacognitive skills and reading. In P. D. Pearson, M. Kamil, R. Barr, & P. Rosenthal (Eds.), *Handbook of reading research* (pp. 353–394). New York: Longman.

Bouffard-Bouchard, T. (1994). Effect of activating conditional knowledge on self-efficacy and comprehension monitoring. *International Journal of Behavioral Development, 17*(3), 577–592.

Brown, J. I., Bennett, F. M., & Hanna, G. (1981). *The Nelson Denny reading test (form E).* Chicago: Riverside Publishing.

Flesch, R. F. (1951). *How to test readability.* New York: Harper.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 702–718.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15*(1), 84–93.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116*(2), 119–136.

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

Glover, J. A. (1989). Improving readers' estimates of learning from text: The role of inserted questions. *Reading Research Quarterly, 28,* 68–75.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology and Aging, 5,* 215–227.

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

Lin, L., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23*(4), 345–391.

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*(4), 609–616.

Maki, R. H., & Serra, M. (1992a). The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(1), 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*(2), 200–210.

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

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

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*(3), 257–260.

Olin, J. T., & Zelinski, E. M. (1997). Age differences in calibration of comprehension. *Educational Gerontology, 23,* 67–77.

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*(3), 435–437.

Weaver, C. A. III. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 214–222.

Weaver, C. A. III, & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition, 23*(1), 12–22.

Weaver, C. A. III, Bryant, D. S., & Burns, K. D. (1995). Comprehension monitoring: Extensions of the Kintsch and van Dijk model. In C. A. Weaver III, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 177–193). Hillsdale, NJ: Erlbaum.

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., Moore, D., & Schultz, N. R. (1993). Young and old adults' ability to use different standards to evaluate understanding. *Journal of Gerontology: Psychological Sciences, 48*(3), P238–P244.