Physics is Harder than Psychology (Or Is It?): Developmental Differences in Calibration of Domain-Specific Texts

Corinne Zimmerman (czimmer@ilstu.edu)
Sarah Gerson (sgerson@umd.edu)
Andrew Monroe (aemonro@ilstu.edu)
Amanda Kearney (akearney_smith@msn.com)
Department of Psychology, Illinois State University, Campus Box 4620
Normal, IL 61790 USA

Abstract

Comprehension monitoring is an important metacognitive skill that is necessary for students to become proficient readers and learners. *Calibration* is one measure of comprehension monitoring and is calculated by comparing actual comprehension (i.e., score on a test) to perceived comprehension. Domain of text was investigated as one source of calibration variability. In two experiments, we provide evidence that the domain of text can influence calibration performance, and this effect varies with the age and educational level of the individual. Younger students (junior high and high school) were relatively well calibrated, but college students demonstrated both underconfidence (e.g., in physics) and overconfidence (e.g., in psychology). That is, although an "illusion of knowing" has previously been reported in the literature, in some domains students showed an "illusion of *not* knowing."

Keywords: calibration; metacomprehension; metacognition; domain specific; cognitive development; psychology

Introduction

Comprehension is defined as the ability to understand and process text. Comprehension monitoring, in contrast, is a metacognitive skill that involves readers' ability to predict which passages they have understood and which they have not. As readers develop and refine metacognitive abilities, the reading process may become more accurate and efficient (Weaver & Bryant, 1995). Several studies have examined the relationship between college students' comprehension abilities and their perceptions of their own abilities (e.g., Glenberg & Epstein, 1985; Morris, 1995; Weaver & Bryant, 1995; Lin, Moore, & Zabrucky, 2001), with the most common finding being one of a mismatch – most often as an "illusion of knowing" (for a review, see Lin & Zabrucky, 1998). In the current educational climate, assessment of comprehension is a common component of standardized testing. Does this same mismatch between perception and actual comprehension exist for younger students who are still developing metacomprehension abilities?

Calibration is a measure of comprehension monitoring that is calculated by comparing actual comprehension (i.e., score on a test) to perceived comprehension (Glenberg & Epstein, 1985). Most calibration studies have been conducted with college students, and researchers have examined numerous participant, text, and task variables that influence calibration. Overall, Lin and Zabrucky (1998) concluded that calibration is generally inaccurate among college students. However, only a few researchers have examined the *development* of calibration. For example, Brubaker-Ward (1995) found a developmental trend in that calibration was better for graduate students than for undergraduates, with both groups performing better than high-school students. Brubaker-Ward hypothesized that better comprehension monitoring develops with education and practice, not just with age. For a more complete picture of the development of comprehension monitoring, it is necessary to evaluate younger populations of readers.

Domain-specific knowledge has been shown to affect comprehension ability (e.g., Recht & Leslie, 1988; Stahl et al., 1989; Walker 1987), so it is reasonable to expect that it may also influence measures comprehension monitoring. The few studies that address the effect of domain-specific knowledge on calibration have shown mixed findings. Both Schraw (1997) and Winne and Muis (2003) found no evidence that calibration is domain-specific, and Schraw concluded that calibration is a general metacognitive ability. In contrast, Glenberg and Epstein (1987) found evidence of the importance of domain by assessing calibration for participants with expertise in either music or physics. They hypothesized that students would calibrate substantially better in their area of expertise than in the other domain. In the domain for which they had no particular expertise, students' actual and predicted performance matched. However, students were poorly calibrated on texts that matched their expertise – there was almost no correlation between actual and predicted performance. Glenberg and Epstein suggested participants' confidence judgments were based on familiarity with the domain rather than information gained from the text.

These mixed results necessitate further study of the role of domain. Domain is a particularly important factor, given that the texts that students encounter and learn from are domain-specific, as are text passages used in standardized

tests of comprehension. Although domain knowledge has been shown to facilitate comprehension, more evidence is needed of the role of domain in calibration ability, because "prior knowledge and domain interest can become counterproductive when students rely too heavily on them to make decisions about when to terminate further text processing" (Lin & Zabrucky, 1998, p. 41).

In the current study, educationally relevant domains and cohort were investigated as potential sources of calibration variability. The domains selected (biology, chemistry, physics, psychology, math and history) appear at every level of schooling with the exception of psychology (high school and college only). Therefore, it is reasonable to assume that students have been exposed to these domains through standard curricula. Our specific aim was to address the following questions: (a) Are there age/education differences in overall calibration?; (b) Does calibration differ by domain?; and (c) If domain differences exist, which domains lead to the most accurate calibration?

Experiment 1

Participants

The sample included 27 junior-high students (*M* age 13.4, *SD* = .70), 38 high-school students (*M* age 17.5, *SD* = .76) and 31 college students (*M* age 20.5, *SD* = 2.1). There were approximately equal numbers of male and female participants and the racial/ethnic make-up was representative of the population of a Midwestern school district.

Materials and Procedure

Participants read text passages from four domains: biology (evolution), chemistry (periodic table), physics (quantum physics), and psychology (classical conditioning), with order of presentation counterbalanced. All texts were expository (*M* = 208 words). Text difficulty was controlled for each age group to maintain grade-appropriate levels using the Flesch-Kincaid Readability Scale (1951; 1973), which produces a reading ease level indicating the degree of complexity based on sentence structure and language difficulty. Text passages were adapted from Tallack (2003) following Glenberg and Epstein's (1985) standards for composing texts: (a) texts should be self-contained and (b) texts should be organized about an explicitly stated central proposition.

Students were tested in groups and instructed to read each passage as though studying for an exam. After reading each passage, individuals provided a confidence judgment indicating their percieved ability to answer questions about the text using a 5-point Likert scale (*1- extremely unconfident* to *5- extremely confident*). Five multiple-choice questions assessed overall comprehension of the material rather than testing specific content. Questions were based on thematic schemes derived from standardized sample exam questions (Martinson, 1993). Students were instructed not to look back at text passages while answering questions.

Results

Overall calibration. Previous researchers (e.g., Glenberg & Epstein, 1985; Weaver, 1990) computed calibration by correlating actual performance with confidence judgments. Gamma correlations (γ) are typically reported (Nelson, 1984, cited in Lin et al., 2001) because of the ordinal nature of the data. The differences in mean gamma correlations for junior-high (M = .25; SD = .69), high-school (M = -.08; SD = .87) and college students (M = .09; SD = .74) were not statistically significant, $F(2, 92) = 1.41$, $p = .25$, $\eta^2 = .030$. The gamma values are typical of those reported in the literature (e.g., Lin & Zabrucky, 1998) and indicate inaccurate calibration at all age levels. However, a different picture emerges when each domain is examined separately.

Calibration as a function of domain and cohort. To determine if students' calibration ability differs depending on the domain of the text passage, it was necessary to use a different measure of calibration because correlations depend on multiple texts. A simple difference score was used. Confidence judgments (i.e., Likert scale) and comprehension scores (i.e., actual number correct) are not the same kinds of measurements, so both scores were *Z*-transformed. A difference score based on *Z*-scores was then computed for each domain. Difference scores (confidence – actual) were subjected to a mixed 3 (cohort) x 4 (domain) ANOVA. The main effect of cohort was not significant, $F(2, 91) = 1.30$, $p = .28$, $\eta_p^2 = .028$. There was a main effect of domain, $F(3, 89) = 12.78$, $p < .001$, $\eta_p^2 = .301$, which was qualified by a significant interaction, $F(6, 178) = 2.93$, $p = .009$, $\eta_p^2 = .087$. As seen in Figure 1, calibration for texts in different domains varied by the educational level of the students.

Accuracy of calibration. For difference scores, values of zero represent accurate calibration (i.e., the confidence judgment and the actual comprehension score was the same). In contrast, values above or below zero represent "overconfidence" or "underconfidence," respectively. One-sample *t*-tests with a test value of zero were used to determine which domains were associated with accurate calibration (alpha was set at 0.0125 to account for multiple tests). For junior-high and high school students, calibration was relatively accurate for the biology, chemistry and physics texts, indicating that confidence for answering questions was similar to actual comprehension performance. Overconfidence was found for the psychology text for all students, suggesting that at all levels students were more confident than their performance warranted. College students were accurate for biology but under-confident for chemistry and physics texts.

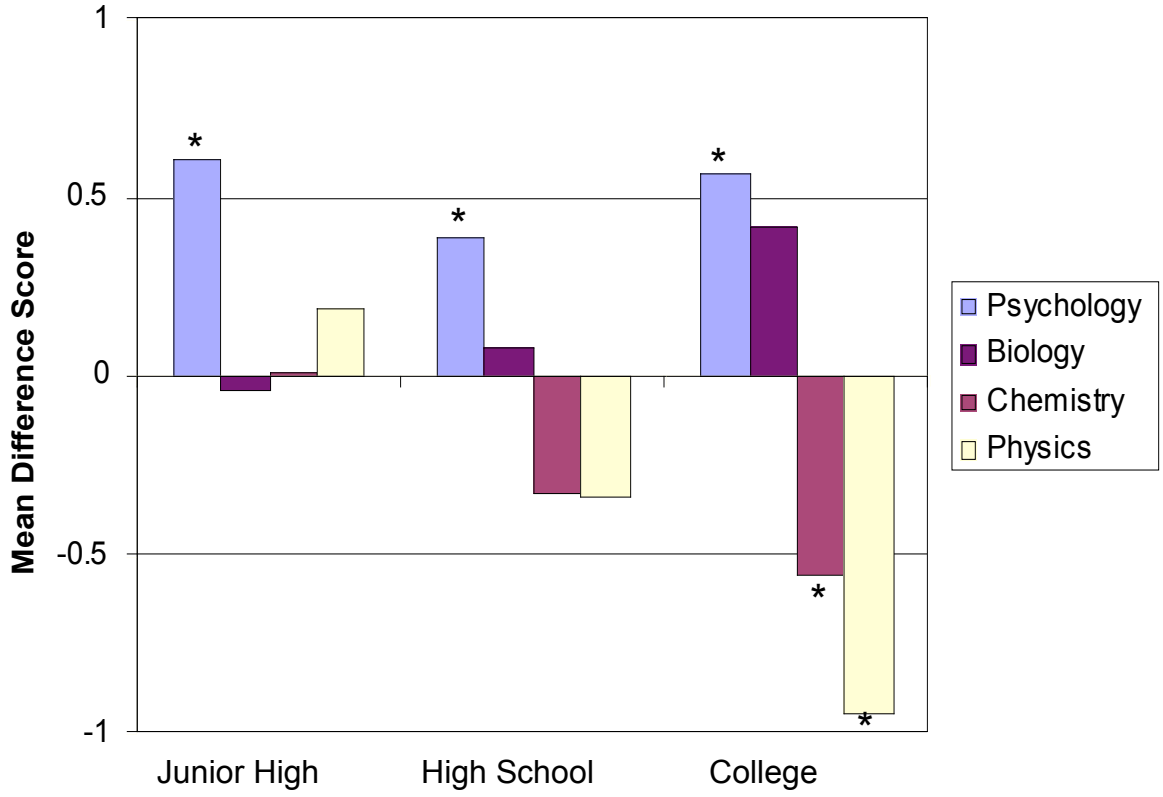

Figure 1. Mean difference scores (confidence judgments – comprehension score) as a function of domain and cohort for Experiment 1. Positive values represents "overconfidence" and negative values represent "underconfidence." Scores significantly different from zero (i.e., accurate calibration) are denoted with an asterisk ($p < .0125$).

Discussion

Interestingly, junior-high students were reasonably well calibrated for traditional science domains. This finding is counter to what would be expected from research on general or specific metacognitive abilities (e.g., Kuhn & Pearsall, 1998; Moshman, 1995; Schraw & Moshman, 1995), which would predict that the youngest students would be inaccurately calibrated. At the high school level, a similar pattern was found, but the pattern in Figure 1 is suggestive of the beginnings of underconfidence for chemistry and physics. By college, students' performance was sometimes underconfident and sometimes overconfident, depending upon domain. This contrasts with the commonly reported inaccuracy in calibration as one of overconfidence – or an "illusion of knowing" (e.g., Glenberg et al., 1982).

Time constraints precluded using more than the four texts (participants completed other tasks in the same session), limiting our ability to generalize the results. Therefore, a second experiment was conducted as a replication and extension. Five domains were selected, with two texts from each domain. Students were asked to provide both confidence judgments and a prediction of the number of questions they would answer correctly. Confidence judgments allow a comparison with Experiment 1, and predictions allow for a second measure of calibration.

Experiment 2

We were specifically interested in knowing: (a) Will the domain x cohort interaction found in the first experiment be replicated? (b) Which domains are under- and overcalibrated for the different developmental levels? (c) Does the same general pattern of results hold when calibration difference scores are computed using predicted correct versus confidence judgments? The general procedure was the same as that used in Experiment 1. Differences are noted below.

Participants

The sample included 25 junior-high students (*M* age 13.7, *SD* = .54), 27 high-school students (*M* age 17.6, *SD* = .49) and 31 college students (*M* age = 19.6, *SD* = 1.1).

Materials and Procedure

Participants read ten text passages (*M* = 308 words) covering topics in five domains: biology (evolution, DNA),

physics (quantum physics, thermodynamics), psychology (classical conditioning, cognitive development), math (pi, zero), and history (political parties, electoral college). In addition to answering comprehension questions and providing a confidence rating, participants were asked to predict the number of questions they would answer correctly after reading each text.

Results

Results are presented for the two different measures of calibration using difference scores. Calibration was first computed using participants' confidence judgments in order to make comparisons with Experiment 1. A second measure of calibration was compared predicted number correct with actual number correct.

Calibration using Z-transformed confidence scores. As in Experiment 1, actual scores and confidence ratings were *Z*-transformed prior to computing difference scores. The main effect of domain was not significant, $F(4, 76) = 1.78$, $p = .14$, $\eta_p^2 = .086$. A significant cohort effect was found, $F(2,79) = 6.79$, $p = .002$, $\eta_p^2 = .147$, which was qualified by a significant interaction, $F(8, 152) = 5.25$, $p < .001$, $\eta_p^2 = .216$. As seen in Figure 2, the pattern for the three repeated domains (psychology, biology and physics) is similar to that found in Experiment 1, with the exception that high school students were not overconfident for psychology. With respect to the two additional domains, all students showed accurate calibration for texts about mathematics. For history texts, junior high students were overconfident but college students were underconfident.

Calibration using predicted and actual scores. Difference scores were computed comparing the predicted number correct with the actual number correct. Positive values indicate overconfidence whereas negative values indicate underconfidence, with numbers near zero indicating accurate calibration. A marginal effect of cohort was found, $F(2, 75) = 3.00$, $p = .056$, $\eta_p^2 = .074$. A main effect of domain was evident, $F(4,72) = 10.4$, $p < .001$, $\eta_p^2 = .366$, as was the cohort x domain interaction, $F(8, 144) = 2.89$, $p = .005$, $\eta_p^2 = .138$ (see Figure 3).

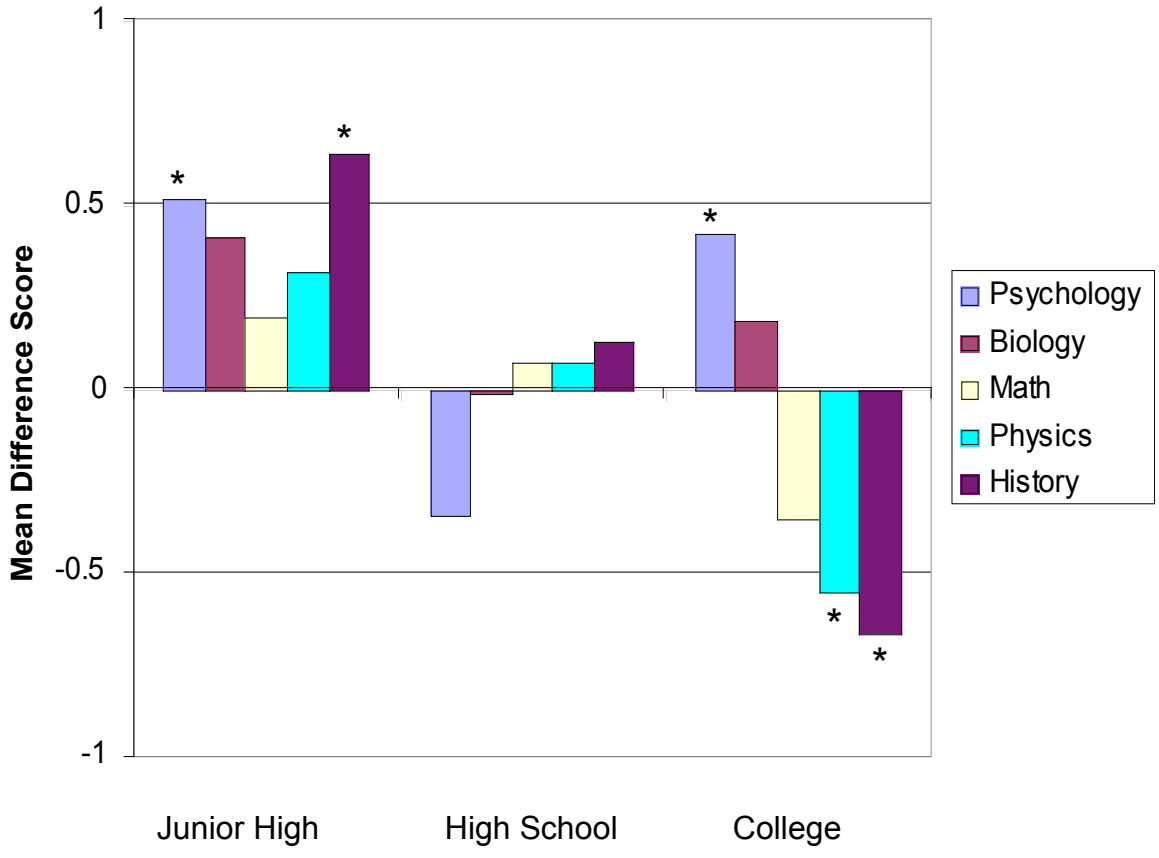

Figure 2. Mean difference score as a function of Domain and Cohort for Experiment 2. Difference scores are computed from *Z*-transformed comprehension scores and confidence ratings. Scores significantly different from zero (i.e., accurate calibration) are denoted with an asterisk ($p < .0125$).

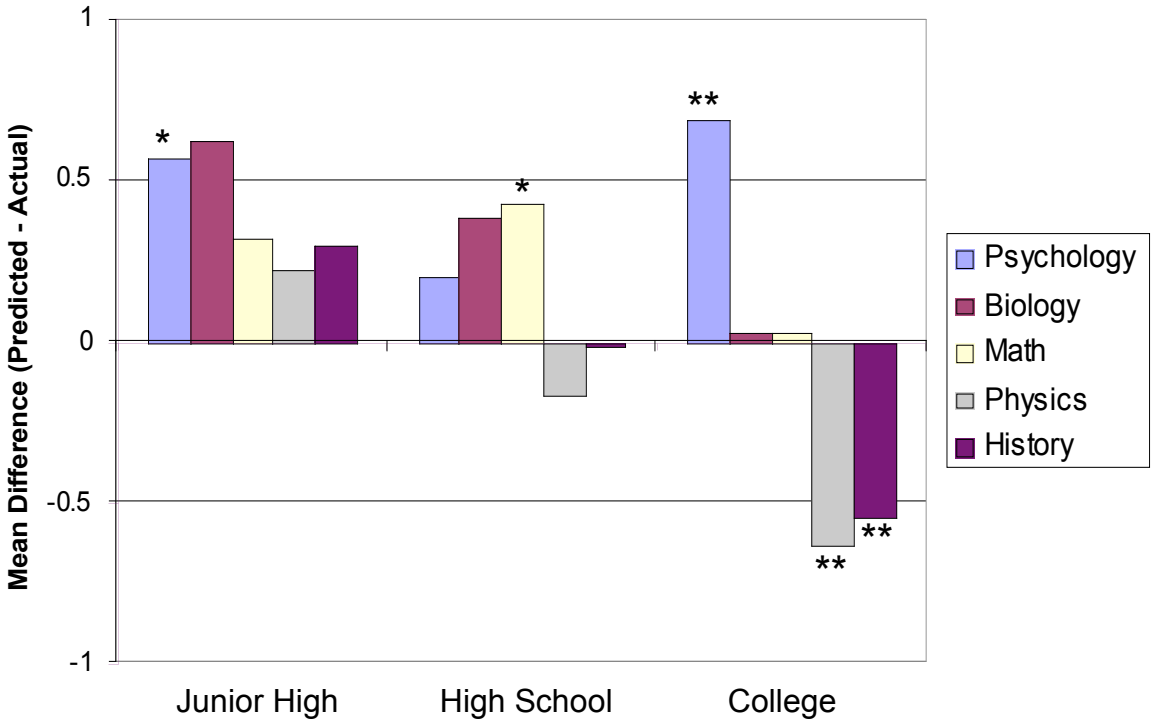

Figure 3. Difference scores computed by subtracting actual number correct from the predicted number correct shown as a function of Domain and Cohort. Scores significantly different from zero (i.e., accurate calibration) are denoted with an asterisk (* $p < 0.0125$; ** $p < .01$).

Discussion

The second experiment served as a replication and extension with the addition of two domains, two texts per domain, and an additional calibration measure. With the exception of accurate calibration of the psychology text for high school students, the pattern for the three repeated domains was similar across experiments when computing calibration with confidence judgments. Confidence in one's ability to answer questions about a text passage, and a prediction about the number correct were highly correlated (all were significant and ranged from .50 to .84). There were similarities for calibration computed using confidence judgments and predicted number correct, with some notable differences (calibration scores for history at the junior high level and for math at the high school level). Overall, the pattern of results was quite similar.

General Discussion

The current study provides evidence that the domain of text can influence calibration performance, and that this effect varies with the age and educational level of the individual. Although the most commonly reported finding is one of overconfidence (the illusion of knowing) in comprehension abilities, in both experiments we also found evidence of an "illusion of *not* knowing" in the domains of physics, chemistry and history. Although this finding was unexpected, there are two possible interpretations of the developmental trend showing increased inaccuracy with increased educational level. First, it could be that as students advance through school, they may develop conceptions about different subject areas that are based on "reputation" of the domain (e.g., "physics is hard"; "psychology is easy") that could interfere with monitoring of comprehension. Second, the underconfidence shown by college students may be a sign of more sophisticated metacomprehension. That is, because they assessed certain texts as "difficult" (e.g., "this passage is about physics so it is going to be difficult") they may have engaged in strategies such as slowing down their reading or re-reading sentences, which would result in the mismatch between confidence and performance. This interpretation, however, requires further research.

Inaccurate calibration, particularly in the form of overconfidence, may indicate that students are not productively monitoring their comprehension (Pintrich,

2002; Winne & Muis, 2003). Comprehension monitoring is an important metacognitive skill for learning in general, and assessment in particular. Standardized testing situations are ones for which the ability to monitor reading comprehension becomes crucial for successful performance. Of particular concern is the situation in which students may unintentionally disregard information presented in the text and answer based on prior knowledge or the reputation of the domain. On the ACT reading test, for example, sample and test questions come from domains of social science, natural science, and prose fiction.[1] In the current educational climate of increased accountability through standardized tests, research aimed at understanding the factors that influence comprehension monitoring is important to prepare students to become proficient readers, learners and test-takers.

Acknowledgements

Sarah Gerson is now a graduate student in the Department of Psychology, University of Maryland. Amanda Kearney is now in the Community Development Division at the Department of Family and Community Services in Albuquerque, NM and teaches part time at the Central New Mexico Community College.

References

Brubaker-Ward, S. (1995). The effects of education level on illusion of knowing and comprehension monitoring activity. *Educational Research Quarterly,* **19***,* 23-41.

Flesch, R. (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**, 702-708.

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

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

Kuhn, D., & Pearsall, S. (1998) Relations between metastrategic knowledge and strategic performance. *Cognitive Development, 13*, 227-247.

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.

Martinson, T. H. (1993). *GRE Supercourse (3rd Ed.)*. New York: Prentice Hall.

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

Moshman, D. (1995). Cognitive development beyond childhood. In D. William (Ed.), *Handbook of Child Development, Volume 2: Cognition, perception, and language*. Hoboken, NJ: John Wiley & Sons.

Pintrich, P. (2002). The role of metacognitive knowledge in learning, teaching and assessing. *Theory into Practice*, **41**, 219-225.

Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers'memory of text. *Journal of Educational Psychology, 80,* 16-20.

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., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review*, **7**, 351-371.

Stahl, S. A., Jacobson, M. G., Davis, C. E., & Davis, R. L. (1989). Prior knowledge and difficult vocabulary in the comprehension of familiar text. *Reading Research Quarterly,* **24***,* 27-43.

Tallack, P. (Ed.). (2003). *The Science Book*. London: Wiedenfeld & Nicolson.

Walker, C. H. (1987). Relative importance of domain knowledge and overall aptitude on acquisition of domain-related information. *Cognition and Instruction,* **4,** 25-42.

Weaver, C. A. III (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* **16**, 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**, 12-22.

Winne, P. H., & Muis, K. R. (2003, April). *Can statistical estimates replace learners' judgments about knowledge in calibration of achievement?* Paper presented at the meeting of the American Educational Research Association Convention, Chicago, IL.

[1] For example, sample text passages and questions for the ACT are available at: http://www.actstudent.org/sampletest/test4/read4/readingtest.html