

Expertise and the Illusion of Comprehension

Ben Jee (bendj@uic.edu), Jennifer Wiley (jwiley@uic.edu)
& Thomas Griffin (tgriffin@uic.edu)

Department of Psychology, 1007 W. Harrison St.
University of Illinois, Chicago
Chicago, IL 60607 USA

Abstract

For almost 20 years the findings of Glenberg and Epstein (1987) have been offered as evidence that experts are less accurate than novices in judging the state of their understanding. In this paper we point to a number of difficulties with this finding, and report a modified replication of the original study. Although we find that the possession of domain knowledge is positively related to both comprehension and judgments of comprehension, domain knowledge was not significantly related to relative metacomprehension accuracy. In terms of absolute accuracy, however, higher-knowledge individuals actually appear better calibrated. Altogether, we find that domain knowledge accounts for an extremely slight proportion of the variance in the accuracy of comprehension judgments.

Expertise and the Illusion of Comprehension

Close to twenty years of research has cited the finding of Glenberg and Epstein (1987) that experts can be less accurate about their own state of understanding than novices. In the original study, Glenberg and Epstein found a negative relationship between the accuracy of metacomprehension judgments (readers' predictions of whether they would be able to answer comprehension questions about texts they had just read) and the number of courses they had taken in the content area. To our knowledge these results have not been replicated, yet they form the basis for over 100 citations of the claim that there is a negative effect of expertise or familiarity with content on comprehension monitoring judgments (c.f. Glenberg, Sanocki, Epstein, & Morris, 1987).

One of the main reasons for the popularity of Glenberg and Epstein's finding is likely that it stands in contrast to most findings related to domain knowledge and performance on domain-related tasks, i.e. that the possession of domain knowledge usually improves performance on domain-related text comprehension and problem solving (Chi, Glaser & Farr, 1988; Ericsson & Kintsch, 1995; Spilich, Vesonder, Chiesi & Voss, 1979). This conclusion suggests that experts may be less likely to review information that they fail to comprehend accurately on their first pass through a novel text. Moreover, if high-knowledge individuals assume that they understand novel domain-related information on the basis of their general proficiency in the domain, they may inaccurately encode this information, or fail to encode it altogether. This may increase the probability that these individuals will commit undesirable and potentially costly errors.

A Closer Look at the Illusion

Despite the popularity of Glenberg and Epstein's (1987) conclusions, there are a number of concerns that can be raised about their study. A principal concern is the way in which Glenberg and Epstein analyzed their data on the relation between expertise (number of courses taken) and the accuracy of metacomprehension judgments. The Goodman-Kruskal Gamma (G) was used as an index of the association between metacomprehension judgments and comprehension scores. Higher values for G imply better *relative* accuracy of comprehension judgments. For each participant in their study, a separate G was computed for each domain of texts presented, music and physics. Separate regression analyses were then conducted to predict the mean G for each domain, with number of music courses (MC) and number of physics courses (PC) serving as the two predictor variables in each equation. These regression equations are reproduced in Table 1.

Table 1: Regression equations for resolution of comprehension (from Glenberg & Epstein, 1987, p.87)

DV	Equation
G_{MUSIC}	$0.10 - 0.025\text{MC} + 0.012\text{PC}$
G_{PHYSICS}	$0.37 - 0.021\text{MC} - 0.117\text{PC}$

Note. G_{MUSIC} = Mean Gamma for music; G_{PHYSICS} = Mean Gamma for physics; MC = Number of music courses; PC = Number of physics courses.

Glenberg and Epstein found a statistically significant difference in the coefficients for PC (+0.012 and -0.117) between the equations for G_{MUSIC} and G_{PHYSICS} . (Interestingly, the difference between the coefficients for MC, -0.025 and -0.021, did not differ between the equations). This significant difference was taken as evidence that expertise (i.e., number of physics courses) is inversely related to accuracy. The issue that we wish to draw attention to is that finding such a difference between two regression equations does not imply that a variable is a significant predictor within either of the equations. The difference in the coefficients for PC between the two regression equations only implies that number of physics courses differs in predicting mean G for music vs. physics, but it does not imply that number of physics courses is a significant negative predictor of G for physics (or that it is

a significant positive predictor of music G_c for that matter). The latter conclusion, however, is the one that is so often cited. To affirm this conclusion would require a test of whether the relation between PC and G_{PHYSICS} is significantly different from zero. Although the coefficient for PC is negative in equation for G_{PHYSICS} , the question remains whether there is a significant negative relation between expertise and metacognitive accuracy.

Besides performing a direct test of the relation between expertise and accuracy, it is also important to consider the degree to which variance in expertise accounts for the variance in accuracy. Glenberg and Epstein fail to report an index of fit for their regression equations, for example, the R^2 statistic. How well do these equations account for their data? If the fit is very poor, then there is little reason to take the relation between expertise and accuracy very seriously. Another concern is that Glenberg and Epstein found positive or null relations between music expertise and the accuracy of comprehension judgments. In particular, the coefficients associated with number of music courses did not differ between domains. Therefore, beyond the insufficient evidence for the existence of the illusion of comprehension in physics, there was no evidence for the illusion in the other domain that they examined.

Certain features of the methodology of the original study also warrant concern. First, the measure of comprehension that was used in computing metacognitive accuracy was a single true/false test item. Reliability problems due to assessing comprehension with a single test question have been discussed in the metacomprehension literature previously (Glenberg, Sanocki, Epstein & Morris, 1987; Maki & Serra, 1992; Weaver, 1990; Wiley, Griffin & Thiede, 2005). In this particular case, where the measure of interest is a correlation between test performance and readers' predictions of their performance, measures with a larger range than zero and one would be preferable. In order to see if readers really can gauge their own understanding of text, more assessment items would allow for a more sensitive test of the relation.

A second methodological issue in the original study is the way expertise was measured. Participants reported the number of courses taken in a domain related to each set of texts (on music and physics). Of course, this yielded a highly skewed distribution with many students taking only one or two courses. Although courses taken is one proxy for expertise, measures of the possession of domain knowledge that might yield a distribution with better range, and potentially more valid scores, for correlational analyses. Altogether, there are several compelling reasons for re-examining the connection between expertise and the accuracy of judgments of metacomprehension.

Overview of Present Study

In the present study, we perform a replication of Glenberg and Epstein's original study, while keeping in mind the

concerns discussed above. Our participants are each presented with a practice trial, 5 baseball-related texts, and 5 general knowledge texts. Each text has 5 short-answer test questions. After reading each text, readers are asked to predict how they would perform on a 5-item test on the text by selecting how many questions that they think they would be able to answer correctly, from 0 to 5. After reading and judging their comprehension for a set of texts, participants are tested on each text in the set. The order of sets is counterbalanced across subjects. Finally, domain knowledge is measured with a 45-item baseball knowledge questionnaire.

If expertise, or increasing domain knowledge, were truly associated with illusions of comprehension, then we would expect a significant negative relation between score on the baseball knowledge questionnaire and metacomprehension accuracy. If domain knowledge, on the other hand, permits readers to monitor their performance and make better assessments of what they know after reading a text (c.f. Glaser & Chi, 1988), then metacomprehension accuracy should increase with domain knowledge.

Method

Participants

114 undergraduates at the University of Illinois at Chicago received course credit for their participation in this experiment as part of an Introductory Psychology subject pool. Participants were treated in accordance with the "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 1992). Data from 13 participants was eliminated because of lack of variability in their judgments of comprehension. Data from three participants were lost due to equipment failure. This left usable data from 98 participants.

Materials

Five texts about baseball, five general knowledge texts, and one practice text were written for this study. The texts about baseball were on bunting strategy, corking bats, a new batting average statistic, the crack of the bat when the ball is hit, and curve balls. The general knowledge texts were on the potato famine, cell division, dinosaur extinction, electric cars, and acquired heart disease. The practice text was on stalactites and stalagmites. The texts were all around 400 words in length. For each text, 5 test items were constructed with the intention that at least one would be relatively easy for novices (a gist question or a question about a specific fact mentioned in the text), while several questions required inferences from each text. We wrote questions that required the reading of the text in order to be answered correctly, even for experts.

The 45-item Baseball Knowledge Questionnaire was taken from Spilich, Vesonder, Chiesi, and Voss (1979). In addition, questions were added about interest in baseball, experience playing and coaching baseball, familiarity with the general knowledge and baseball-related topics, and the number of physics courses taken.

Procedure

Participants were tested in small groups (≤ 10). Each participant was seated at a desk in front of a computer, and the task instructions, texts, judgment items, and test items were presented on computer. The Baseball Knowledge Questionnaire was administered as a paper-and-pencil booklet. In the general instructions, the participant was informed that they would be asked to read several texts, and that they should read each text carefully as if studying for an exam. They were also told that they could read each text at their own pace and that rereading was allowed, but once they finished a text and advanced the screen they could not go back. Finally, participants were informed that they will have to make a judgment about the number of questions (0-5) that they think they will be able to answer about each text, and that they will have to answer several questions about the texts after reading them.

After reading the general instructions, participants were given a practice text and test items. The entire practice text was presented on a single screen. At the bottom of the screen there was a link to another screen where the judgment item was presented. After reading the text the participant could click this link using the computer mouse. On the judgment item screen the participant was asked to judge how many questions (0-5) they think they will be able to answer about the practice text. The participant could click on a given value using the computer mouse. After making their judgment, they could use the mouse to click an onscreen button to advance to the next screen. The following five screens presented questions about the practice text. On each of these screens a question appeared above an empty box in which the participant could type their response using the computer keyboard. After typing their response they could submit their answer and advance to the next question by clicking an onscreen button.

After the practice, participants were given further instructions about the actual task. They were informed that they would have to read and make judgments for five texts in a row, and that they would be given a number of questions about each text after reading all five texts. They were also informed that they would be presented with second block of five texts after the first. For each block of five texts, the participant was presented with a judgment screen (as in practice) immediately after reading each text. After reading and making judgments for all five texts, the participant completed five test items on each text in the order of text presentation. The baseball-related and general knowledge texts were presented in different blocks. Also, we created two different text orders for each block. Block order and text order was counterbalanced approximately equally across participants. After completing both blocks of texts, the participant completed the paper-and-pencil Baseball Knowledge Questionnaire with the added questions. The entire procedure took about 80-90 minutes.

Results

Baseball Knowledge

The mean score on the Baseball Knowledge Questionnaire was 15.0 out of 45 ($SD = 12.7$). Table 2 displays the frequencies for different score intervals. From this table it is clear that the distribution of scores is positively skewed. The modal score on the questionnaire is between 0 and 5, however, several individuals performed quite well, thus driving up the mean score. Overall, a fairly wide range of baseball knowledge is observed in our sample.

Table 2: Grouped frequency table for baseball knowledge test scores

Score	Frequency
0-5	34
6-10	14
11-15	6
16-20	10
21-25	11
26-30	6
31-35	9
36-40	7
41-45	1

Note. Maximum score is 45.

Text Comprehension

On average, participants answered 3.1 out of 5.0 questions correctly for each of the five baseball texts in the set ($SD = 0.82$). Figure 1 plots average test performance as a function of score on the Baseball Knowledge Questionnaire. As is clear from the figure, participants with more baseball knowledge performed better on the test questions ($r(98) = 0.61, p < .01$). Importantly, even individuals who scored especially poorly on the Baseball Knowledge Questionnaire were able to answer about 2 or 3 questions correctly, thus avoiding floor effects. Furthermore, individuals with the highest scores on the Questionnaire did not reach ceiling levels of performance. Overall, participants with more baseball knowledge are found to perform better on text-related questions. This finding aligns with much prior research demonstrating the positive relation between domain knowledge and text comprehension.

Judgments of Text Comprehension

For the baseball-related texts, participants judged that they would be able to correctly answer on average 2.5 out of 5.0 test questions per text ($SD = 1.13$). Figure 2 plots average comprehension judgments as a function of score on the Baseball Knowledge Questionnaire.

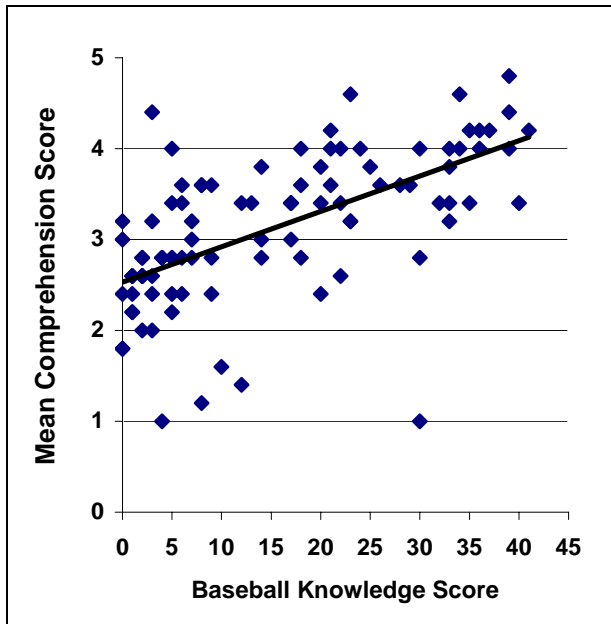


Figure 1: Mean test scores for the baseball texts as a function of Baseball Knowledge Score.

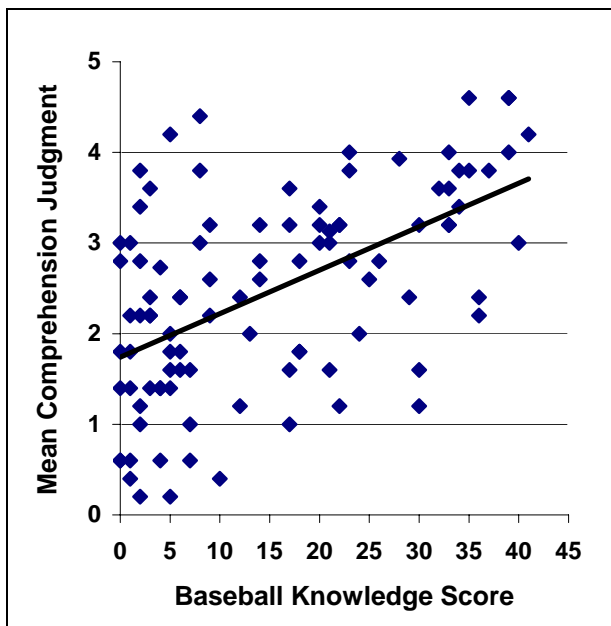


Figure 2: Mean comprehension judgments for the baseball texts as a function of Baseball Knowledge Score.

As displayed in Figure 2, participants with more baseball knowledge judged that they would be able to correctly answer more questions on the baseball texts ($r(98) = 0.54, p < .01$). This is an intuitive finding, and is consistent with Glenberg and Epstein's finding that confidence increases with domain knowledge in the test domain.

Metacomprehension Accuracy

The main question of this paper concerns the relation between participants' domain knowledge and their ability to predict their comprehension of domain-related texts.

We have already observed that both test performance and judgments of text comprehension increase with domain knowledge. The present question is the how the relation between participants' judgments of comprehension and their actual performance varies as a function of domain knowledge.

Metacomprehension accuracy is the relation between predictive judgments and actual performance on the comprehension tests and can be assessed in either relative or absolute terms. In Glenberg and Epstein's original study, *relative* metacomprehension accuracy was analyzed using the Goodman-Kruskal Gamma (G) statistic. Because G is most appropriate for categorical measures, and we have continuous data for both comprehension measures and judgments, we report Pearson correlations. In all of our analyses, however, both of these measures led to convergent results. In using Pearson correlations, our first concern is with the individual variability of the factors involved. If there were a restricted range in our participants' comprehension scores, comprehension judgments or baseball knowledge scores, this would compromise our analyses. As we have already discussed, however, a great deal of variability is observed for each variable of interest.

For the baseball-related texts, mean relative metacomprehension accuracy was 0.18 (SD = 0.53), which was significantly greater than zero, $t(97) = 3.37, p < .01$. For comparison's sake, this corresponds to a Gamma of 0.20 (SD = 0.75), which is greater than zero, $t(97) = 2.62, p < .05$. This value of G exceeds the values observed by Glenberg and Epstein for the domains of music ($G = 0.06$) and physics ($G = 0.02$), which were not significantly greater than zero. For further comparison, Maki (1998) reported an average G of 0.27 across a number of studies. Thus, unlike the findings of Glenberg and Epstein, our result is a fairly typical value of relative metacomprehension accuracy.

Now let us turn to the main question of this paper: Does metacomprehension accuracy decrease with greater domain knowledge, as Glenberg and Epstein originally concluded? To address this question we computed the correlation between relative accuracy and score on the Baseball Knowledge Questionnaire. Note that if we were to perform a regression analysis using baseball knowledge to predict relative accuracy, the value of the coefficient for baseball knowledge would be equivalent to the correlation between these two variables. Our question is whether the correlation between baseball knowledge and relative accuracy is significantly different than zero.

Figure 3 displays relative accuracy as a function of score on the Baseball Knowledge Questionnaire. We see that relative accuracy does decrease with increasing baseball knowledge. The magnitude of this relation, however, although in the negative direction, is extremely slight and is not significantly different from zero, $r(98) = -0.076, ns$. From Figure 3 it appears that approximately equal numbers of higher-knowledge individuals have positive

and negative scores. Is the relation between domain knowledge and resolution a meaningful one? Our correlation implies that that variation in expertise accounts for less than 0.1% of the variance in relative accuracy. Thus, in our study at least, domain knowledge is clearly a poor predictor of relative metacomprehension accuracy.

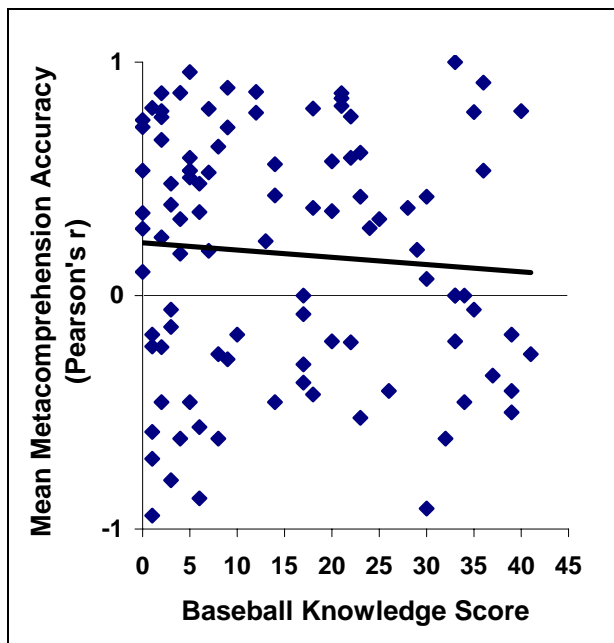


Figure 3: Relative metacomprehension accuracy (Pearson's r) as a function of Baseball Knowledge Score.

As noted above, Gamma and Pearson's r provide indices of *relative* metacomprehension accuracy. Neither index, however, takes into account the magnitude of the deviations between participants' judgments and their actual comprehension scores; that is, *absolute* accuracy, or *calibration*. Thus, as another index of participants' metacomprehension accuracy, we examined the average absolute deviations between their comprehension judgments and their test scores. This involved calculating the absolute deviations between participants' judgments and comprehension scores for each text, taking the average of these absolute deviations, and then subtracting this mean value from zero. With this index, participants whose judgments perfectly matched their comprehension scores would have a score of zero. Participants with increasingly larger discrepancies between their judgments and comprehension scores would have increasingly *lower* scores. Do these deviations vary as a function of domain knowledge? If participants with increasing knowledge are *less* accurate in their judgments of comprehension, then the relation between absolute deviations and domain knowledge should be *negative*. If participants with increasing knowledge are *more* accurate in their judgments of comprehension, then this relation should be *positive*. Figure 4 displays participants' absolute deviations as a function of baseball knowledge. It is clear from the figure that domain knowledge is positively related to the average

absolute deviations between judgments and test score; that is, using deviation scores as an index of absolute metacomprehension accuracy, individuals with higher knowledge are more accurate in their judgments of comprehension. In fact, this positive relation is statistically significant, $r(98) = .25, p < .05$.

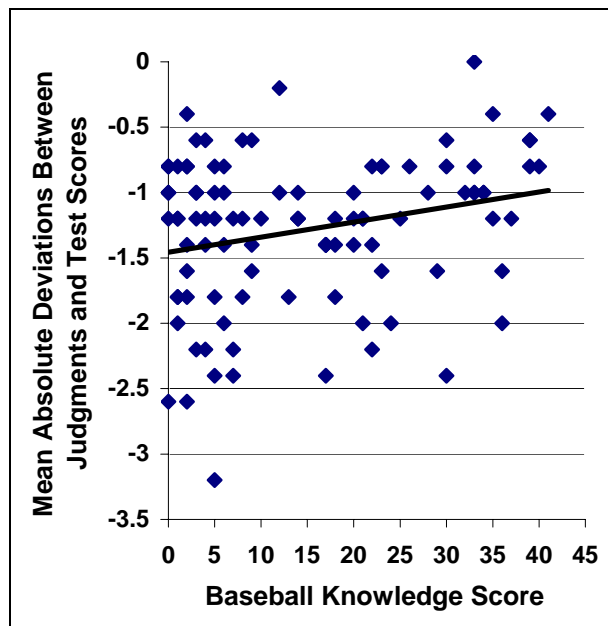


Figure 4: Mean absolute deviations between judgments and comprehension test scores as a function of Baseball Knowledge. *Note.* Mean deviations are subtracted from zero.

Discussion

In terms of *relative* measures of metacomprehension accuracy, such as Gamma or Pearson's r , our study found no significant relation between metacomprehension accuracy and the possession of domain knowledge. Moreover, as we argued above, there are several reasons to doubt that the results of previous research actually support the conclusion of a significant negative relation between metacomprehension accuracy and expertise. In terms of the *absolute* accuracy of participants' judgments, our study did reveal a relation between domain knowledge and metacomprehension accuracy. This relation, however, was in the *positive* direction. Higher-knowledge individuals were actually more accurate in their comprehension judgments. Altogether, as far as we are aware, there is no convincing evidence that experts suffer from the illusion of comprehension that has been repeatedly cited in the literature. If anything, our results suggest that their comprehension judgments may be *more* accurate than lower-knowledge individuals.

Why should we be confident in our result? For one, our sample size was adequate for the correlation analysis that we performed. In fact, Glenberg and Epstein's original analyses of metacomprehension accuracy included only 50 participants, compared to 98 in the present study. The

present study also presented participants with five short-answer comprehension questions for each text, rather than a single true-false item, enabling us to obtain a more sensitive measure of their comprehension. Finally, no ceiling or floor effects were observed with any of the variables that we measured. In contrast, we observed considerable variability. Overall then, there are a number of reasons to be confident in the obtained result of this study. We do, however, plan to recruit more high-knowledge individuals for the present study, as the majority of our sample had low-to-moderate scores on the Baseball Knowledge Questionnaire.

To be clear, we are not concluding that a negative (or positive) relation *cannot* exist between expertise and metacomprehension accuracy. This may very well be the case. For the domain of baseball, however, the present study suggests that this relation may actually be in the opposite direction. Furthermore, there is reason to believe that our finding may be generalizable to other domains. We believe that baseball is a good domain for examining the effects of knowledge, since baseball knowledge lacks an obvious relation to intelligence and reading ability. With other domains, especially academic disciplines, high knowledge may be confounded with both general intelligence and reading ability, making it difficult to assess the unique contribution of domain knowledge.

Conclusions

The original findings of Glenberg and Epstein (1987) have received considerable attention, perhaps because, unlike a wealth of other studies, they highlight a potential pitfall of expertise. In this paper we argued against the validity of these original findings, and reported a study that finds that (a) domain knowledge was a very poor predictor of metacomprehension accuracy using relative measures of accuracy, and that (b) domain knowledge was a positive predictor of accuracy in terms of absolute differences between judgments and comprehension scores.

Although we found a significant positive relation between domain knowledge and absolute accuracy in this study, domain knowledge still did not account for a large amount of the variation in absolute accuracy. From an individual differences perspective, it is still an open and very interesting question why some individuals have almost perfect negative accuracy and others have almost perfect positive accuracy in judging their own comprehension. So, while informative, the results of the present study suggest that factors besides domain knowledge may be more useful in exploring this issue.

Another avenue for future research concerns the effects of domain knowledge in other judgment tasks. Some research has shown that moderate levels of expertise may increase the overall accuracy of probability judgments for domain-related facts (Lichtenstein & Fischhoff, 1977). Other research has found that judgments of explanatory knowledge are generally *overconfident* (Rozenblit & Keil, 2002); how would expertise affect this overconfidence?

Ideally, future research would utilize a single domain of expertise, a single sample of participants, and examine performance across a series of judgment tasks. This would allow a better understanding of how these judgments relate to one another, as well as how they vary both within and across individuals.

Acknowledgements

The authors thank Keith Thiede, Melinda Jensen, Gregory Colflesh, Pat Cushen, Travis Ricks, Christopher Sanchez, and James Voss for their comments and assistance. This research was supported by a grant from the Institute for Education Sciences Cognition and Student Learning program to Keith Thiede and Jennifer Wiley. Any opinions or conclusions expressed here are those of the authors and do not necessarily reflect the funding agency.

References

- Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum.
- Ericsson, KA, & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2), 211-245.
- Glaser, R. and Chi, M. (1988). Overview. In M. Chi, R. Glaser, & M. Farr (Eds.), *The Nature of Expertise* (pp. xv-xxvii). Hillsdale, NJ: Erlbaum.
- 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, 119-136.
- Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior & Human Decision Processes*, 20(2), 159-183.
- 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). Hillsdale, NJ: LEA.
- Maki, R. H., & Serra, M. (1992). Role of practice tests in the accuracy of test predictions on text material. *Journal of Educational Psychology*, 84, 200-210.
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 26, 521-562.
- Spilich, G.J., Vesonder, G.T., Chiesi, H.L., & Voss, J.F. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, 18, 275-290.
- Weaver, C. A. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 16, 214-222.
- Wiley, J., Griffin, T. D. & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *Journal of General Psychology*, 132(4), 408-428.