Metacognition Learning (2006) 1: 159–179
DOI 10.1007/s10409-006-9595-6

The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy

John L. Nietfeld · Li Cao · Jason W. Osborne

Received: 8 October 2005 / Revised: 23 June 2006 /
Accepted: 27 June 2006 / Pulished online: 4 August 2006

Abstract Monitoring one's own study processes accurately is important in self-regulated learning. This study compared a treatment (N = 45) and comparison class (N = 39) on the effects of monitoring exercises and feedback on calibration and test performance over a 16-week undergraduate course. Path analyses revealed a significant influence of the intervention on class performance, calibration, and self-efficacy. The results suggest the appropriateness of integrating distributed metacognitive exercises in class content and the fundamental role of monitoring ability in performance-based course outcomes and self-efficacy.

Keywords monitoring · metacognition

Metacognitive monitoring skills are a core component within information processing models of self-regulation (Butler & Winne, 1995; Winne, 2001) and the development of human expertise in general (Glaser & Chi, 1988). More accurate monitoring has been shown to lead to improved self-regulation that, in turn, translates into improved performance (Thiede, Anderson, & Therriault, 2003). While accurate monitoring by students is certainly not a given (Glenberg & Epstein, 1985; Pressley, Ghatala, Woloshyn, & Pirie, 1990), metacognitive abilities in general are considered to be malleable and largely independent of general intelligence (Pressley & Ghatala, 1990).

Given the importance of monitoring in the self-regulated learning process, researchers now face the challenge of devising ways to increase monitoring accuracy in actual classroom practice (McCormick, 2003). Interventions with elementary students have shown impressive gains in monitoring strategy selection within

J. L. Nietfeld (✉) · J. W. Osborne
North Carolina State University, Curriculum & Instruction,
602D Poe Hall, Raleigh, NC 27695, USA
e-mail: john_nietfeld@ncsu.edu

L. Cao
The University of West Georgia, Maple St., Carrollton, GA, USA

instructional settings that have an explicit focus upon improving monitoring skills (Ghatala, Levin, Pressley, & Goodwin, 1986). Yet, to date there have been few attempts to integrate instructional strategies with the explicit purpose of increasing metacognitive skills for adults over an extended period of time within a classroom context. Therefore, the purpose of the present study was to 1) examine the effects that distributing monitoring exercises and feedback over time have upon performance, the calibration of confidence with performance, and self-efficacy in an actual classroom context and 2) investigate the role that calibration and changes in calibration have upon important course-related outcomes such as test performance, an integrative course project, and self-efficacy.

Defining monitoring accuracy

The regulation dimension of metacognition is generally viewed as effortful, motivated control of one's cognition, memory, or learning that can be subdivided into three different components: planning, monitoring, and evaluation (Schraw & Moshman, 1995). The present study focused on monitoring, which refers to as one's on-line awareness of comprehension and task performance (Schraw & Moshman, 1995). Monitoring is typically viewed as the data-driven dimension of metacognition which functions to inform control processes (Nelson & Narens, 1994). Examples of monitoring are varied but might include judging one's performance on a completed math problem or deciding how well a newspaper article has been comprehended.

The indexes chosen to measure monitoring judgments can be classified as measuring either relative accuracy or absolute accuracy. Relative accuracy is the accuracy of predicting performance on one item relative to another item (Nelson, 1996) whereas absolute accuracy is the extent to which an individual is calibrated with regard to their monitoring judgment and the criterion task (Schraw, 1995). The choice of either a relative or absolute measure of monitoring accuracy should be based upon the context and primary goals of the study (Nietfeld, Enders, & Schraw, 2006). For instance, if the researcher is primarily interested in the extent to which students make consistent judgments across items or can discriminate between materials within a set that lead to poorer performance from materials that lead to higher performance, a measure of relative accuracy may be appropriate (Maki, Shields, Wheeler, & Zacchilli, 2005). On the other hand, if the researcher is interested in comparing judgments with actual performance or with changes in monitoring accuracy related to an intervention, training, or practice effects, a measure of absolute accuracy may be appropriate. Moreover, research has shown measures of absolute accuracy to be more sensitive to individual differences in ability and task variations (Maki et al., 2005). In addition, other constraints related to available sample size, difficulty of the criterion task, and expectation for normally distributed data also factor into one's choice of index (Nietfeld et al., 2006).

In this study we examined measures of absolute accuracy derived from confidence judgments (Schraw, 1994). Confidence judgments have been used to measure comprehension of text, performance on learning tasks, or predictions about future performance. Traditionally, this has been done using Likert scales but recently some have gone to more continuous measures of rating scales (Nietfeld & Schraw, 2002; Schraw, Potenza, & Nebelsick-Gullet, 1993). Continuous scales can provide more precise measurements and are more sensitive to differences that exist

between ratings. Confidence judgments are compared to actual performance and can be used to derive both a measure of calibration and bias either at the local (item-level) or global (test-level) level (Schraw, 1994). Calibration is the process of matching *perception* of performance with *actual* level of performance. When one makes confidence judgments regarding their level of performance, their calibration increases as the absolute difference between their judgments and actual performance decreases. Bias is the extent to which one is over or underconfident. Thus, bias takes into account directionality in the mismatch between judgment and actual performance whereas the focus of calibration is in the absolute difference between the two. This study examined the extent to which college students were calibrated or biased on objective classroom tests and weekly review questions at the local item level and the global level when students were asked to make retrospective global confidence judgments (GCJs) regarding their understanding of class material at the end of each class session.

Previous research in monitoring accuracy

Thus far, research in monitoring has primarily been limited to studying the relationship between monitoring judgments and performance outcomes on objective assessments of recently studied material (for reviews of this literature, see Dunlosky, 2004; Metcalfe, 2000; and Schneider & Pressley, 1997, Chapter 6). Presently, there is little evidence to show how monitoring ability and improvements in monitoring accuracy over an extended period of time relate to complex representations of class material. Moreover, little work has directly examined the relationship between changes in monitoring accuracy and important affective outcomes such as self-efficacy that are essential to the process of self-regulation.

One important exception is provided by Hacker, Bol, Horgan, and Rakow (2000), who also point to the lack of studies on monitoring accuracy in the actual classroom context particularly over an extended period of time. To address this gap, Hacker et al. observed students' global calibration judgments before and after three classroom tests over the semester. Students were encouraged to view the process of self-assessment as a fundamental process to learning and were provided practice tests one week before each test. Results revealed that students' global calibration judgments became more accurate over time, but these improvements were mostly restricted to the highest scoring students on judgments made after the tests. Changes in monitoring accuracy were attributed to the feedback internally generated and externally provided on test performance and the process of making judgments distributed across time. While Hacker et al.'s study was one of the first to focus on criterion outcomes that were meaningful to the student with an emphasis on ecological validity, a) there was no formal intervention with the explicit purpose of increasing monitoring accuracy and b) these authors did not examine changes in self-efficacy, which is a focus of the present research.

The Hacker et al. (2000) findings suggest the need for interventions focused on conditional knowledge and specific procedures to help students to calibrate their perceived understanding of course material with their actual performance. Schunk and Zimmerman (2003) also suggest intentional and systematic interventions as an appropriate approach to develop of self-regulation and components involved in self-regulated learning such as monitoring.

Monitoring accuracy and self-efficacy

Self-efficacy refers to judgment of one's ability to perform a task within a specific domain (Bandura, 1997). Studies have shown the importance of not only considering the ability level of an individual but the individual's belief that they will succeed on a task. Self-efficacy has been shown to contribute a direct effect to performance that is equivalent to general ability itself (Pajares, 1996). Having higher self-efficacy relates to elevated levels of persistence and effort (Bandura & Cervone, 1986; Schunk, 1995) and the tendency to engage in activities to gain new skills (Bandura, 1997).

While the research on the importance of self-efficacy to learning is clear, its relationship with monitoring ability is less clear. Pajares and Kranzler (1995) examined high-school students' judgments of self-efficacy, in the form of confidence judgments, on an item-by-item basis before they solved math problems. Based on the finding that most of the students displayed overconfidence, Pajares and Kranzler suggest the development of instructional interventions to increase students' ability to calibrate their confidence with their performance. Pintrich and De Groot (1990) found that seventh graders with high self-efficacy in science and English courses were more likely to report use of cognitive and metacognitive strategies for such classes. Yet, research to date has not explored how changes in monitoring accuracy over the course of the semester impact changes in self-efficacy using measures taken at the beginning and end of the course. Given that monitoring functions to inform other regulatory processes such as planning and evaluation, is it possible that monitoring also informs self-efficacy in a similar fashion?

The present study

The present study addressed three limitations with current studies of monitoring. First, the study contributed to the literature by implementing an intervention that explicitly focuses on improving monitoring accuracy through distributed exercises and feedback. Past attempts of such interventions for college students have been very limited. Second, little research has tracked such effects over extended periods of time in an actual classroom setting (for an exception see Hacker et al., 2000). Most studies involve one-shot approaches to gathering monitoring judgments. Third, conclusions regarding changes in calibration in relation to performance outcomes have typically been restricted to scores on objective tests, often multiple-choice tests. Little is known about how changes in calibration relate to performance assessment and self-efficacy. For instance, does improved calibration on multiple-choice tests of course content relate to the ability to create more sophisticated representations of knowledge within that domain? Moreover, do gains in calibration correspond with affective gains in confidence to approach challenging problems in the given domain?

Specifically, we addressed three research questions: 1) Do distributed monitoring exercises and feedback improve calibration? 2) Do distributed monitoring exercises and feedback improve student performance outcomes (e.g., test scores, integrative classroom project)? 3) Does calibration or changes in calibration account for the observed changes in performance measures and self-efficacy over the semester?

In order to address these questions, we implemented an intervention that involved two classes of college students in an educational psychology course who provided confidence judgments for each of four multiple-choice classroom tests over the semester. The treatment class completed short monitoring exercises at the end of each class that asked them to assess the extent of their learning for the current class session (GCJ) and their

study preparation. They were also provided with review items in which they provided answers and confidence judgments. Then, answers to the review items were discussed and students were encouraged to reflect on the accuracy of their responses in relation to their confidence judgments before leaving class. In addition, students in the treatment class received feedback on their calibration and bias to monitor accuracy in judging their test performance after each test. The comparison class provided monitoring judgments on each test but did not receive any feedback on the accuracy of their judgments. Students in both classes completed an educational psychology self-efficacy inventory at the beginning and end of the course and completed an integrative classroom project.

Based upon previous studies in monitoring and strategy training (Delclos & Harrington, 1991; Hacker et al., 2000; Nietfeld & Schraw, 2002; Paris, Cross, & Lipson, 1984; West & Stanovich, 1997), we predicted that students in the intervention class would show gradual improvements over that of the comparison group on both calibration and performance for objective tests and on the integrative classroom project. We also hypothesized that the initial measures of test performance would show direct effects with the summative measure of performance and that pretest self-efficacy would show a direct effect for post-test self-efficacy. More importantly, we predicted that there would be indirect paths from the initial measures of test performance through either (or both) average calibration (across tests) and change in calibration to the summative measure of test performance, the integrative course project, and the end of course measure of self-efficacy.

Since there is little existing evidence regarding the relationship between classroom-based monitoring exercises and self-efficacy, we based our predictions upon the prevalent information processing models of self-regulation (Butler & Winne, 1995; Winne & Hadwin, 1998; Winne, 2001). Within these models monitoring functions to drive the self-regulation process by reducing discrepancies between performance standards set by the individual and the goals, tactics, and plans set to achieve those standards. The feedback that is attained over time should allow students to calibrate more accurately and, therefore, create a heightened sense of self-efficacy for learning in the given domain.

Materials and methods

Participants

A total of 84 undergraduate students (85% female) from two sections of an educational psychology survey course voluntarily participated in the study to fulfill a research component for the course. The sections were randomly assigned to the treatment condition (N = 45) and the comparison condition (N = 39). The course was taken during the junior or senior year after admission to the teacher education program at a moderate-sized university in the southeastern United States.

Measures

Four types of outcomes were measured in the study. They included: (a) monitoring accuracy scores, (b) performance scores from objective tests, (c) scores from an integrative project called a schema representation, and (d) self-efficacy scores. Below is a description of these measures in addition to a description of how our, monitoring accuracy indexes were calculated and how monitoring exercises and feedback were provided.

Educational psychology pre-test

A pre-test of educational psychology was given in order to account for background knowledge coming into the course. The test consisted of 25 four-option multiple-choice questions and was not timed. The items covered the range of topics from the course including cognitive and behavioral theories of learning, behavior management, motivation, instructional strategies, and assessment.

Educational psychology self-efficacy

An educational psychology self-efficacy inventory was given during the first class and again during the final class of the semester (see Appendix A). The inventory consisted of eight items answered on a five-point Likert scale. Scores were summed to create one composite score for each administration of the inventory. The self-efficacy scale proved to be internally reliable as evidenced by a coefficient alpha of 0.88 for the pretest and 0.90 for the post-test.

Raven's Advanced Progressive Matrices

Participants completed a portion of the Raven Advanced Progressive Matrices Test (Set II) (Raven, 1962) as a measure of general ability (Carpenter, Just, & Shell, 1990) to determine if there were initial differences between groups. Each problem on the Raven test consists of a 3×3 matrix, in which the lower right entry is missing and must be selected from among eight alternatives. We used an abbreviated version of the 36-item test consisting of problems 6 through 20 that ranged in difficulty from 61 to 90% correct completion.

Test performance

Test performance was measured using the number of test items answered correctly on each of four tests. The first three tests consisted of 20 four-option multiple-choice items while the fourth test was a 40-item comprehensive exam. Each of the first three tests covered a unit of the course content while the final exam was a comprehensive measure of all the content covered in the course. The items were a combination of questions created by the instructors and by those taken from various textbook item banks. They varied in difficulty from simple identification to more difficult application questions. An example of an identification question was:

What type of memory refers to "knowing how" to do something:

A. Episodic
B. Declarative
C. Long term
D. Procedural

An example of an applied question was:

Perry and Louis have just failed at their first attempt at writing an elaborate and intricate computer program that will be used for elementary school students to practice their spelling. Perry, who tends to be oriented towards an internal locus of control, has higher self-efficacy with computer work than Louis who tends to attribute

his successes and failures externally. Given this information what is most likely to happen concerning another attempt at writing the program?

A. Louis will persist in his attempt to write the program but Perry will solve the problem much more quickly
B. Both Louis and Perry will give up as Louis attributes failure to his lack of ability while Perry blames failure on the difficulty of the problem
C. Perry will persist in trying to write the program while Louis moves on to other tasks
D. Perry will blame the failure on bad luck and Louis will claim that they did not expend enough effort

Performance for each item was scored as a 1 if correct and a 0 if incorrect. As each of the four tests produced different means and standard deviations due to the different material being covered, all test and monitoring accuracy scores were converted to z-scores to allow for direct comparison and more substantive interpretation.

For path model analyses our baseline measure of performance was test performance on the first test during the semester. This was the first assessment of knowledge wherein students had the opportunity to study and prepare for a typical course-like test. The summative measure of performance was students' performance on the comprehensive final exam.

Schema representation

A performance assessment called the schema representation project was also used to measure students' performance. All students in the study submitted the project during the final week of the course. The project required students to create a holistic visual representation of what they believed to be the most important components for making them an effective teacher (Nietfeld, 2002). Critical components of the project include providing important course concepts and the justification for including those concepts, integrating knowledge from the course including drawing explicit connections between concepts from different units, and including personal examples of how concepts and theories could be applied in their future classroom. In sum, the goal of the project was for the students to devise a mental model for effective instruction. The final project was completed on a poster board. The project allowed for much creativity in that students could present the information in their mental model by using any metaphor or theme that they chose (e.g., a tree, train, garden, etc.).

The project was scored using five 20-point scales for a possible total score of 100 points. The five scales included visual representation, knowledge assimilation, personalization, accuracy, and effort. A rubric was used in the scoring that included score ranges for performance under each of the five scales.[1] Schema scores were highly skewed (skew < -1.0), and therefore the variable was reflected, a constant was added to anchor the range at 1.0, and the square root was taken. The scores were then returned to their original direction (following the procedure suggested in Osborne, 2002). The resulting potential range of scores for the schema representation was 0 to 8 following this transformation.

[1] Inter-rater reliability was not conducted on the schema representation projects. Instead, a detailed scoring rubric was used to assess the score under each of the five major categories.

Confidence judgments

Participants recorded confidence ratings for each test item on the test of background knowledge, the Raven's Test, and the four classroom tests. The confidence judgments were recorded on a 100-mm line that followed each question (Schiffman, Reynolds, & Young, 1981). The left end of the line corresponded to no confidence and was labeled *0% Confidence*; the right end corresponded to total confidence and was labeled *100% Confidence*. To indicate how much confidence individuals had in their test responses, we asked them to draw a slash through the portion of the line that best corresponded to their perceived confidence on the preceding question. This local measure denotes monitoring that is performed on-line.

Monitoring accuracy

Monitoring accuracy was measured with two indices described by Keren (1991) and Yates (1990). The first index was calibration, which consisted of the absolute value of the difference between the confidence judgment and performance for each test item, summed over all items on a test and divided by the total number of items. Scores could range from zero (perfect calibration) to one (complete lack of calibration). For example, if a confidence rating for a given item is 92 and the participant answered the question correctly the accuracy score for that item would be 0.08 (absolute value of 1–0.92). The second index was the bias score, which consisted of the signed difference between the average confidence and average performance scores on each test. Positive scores indicate overconfidence while negative scores indicate underconfidence. The farther the score is from zero the more biased it is. For example, if a participant has an average confidence rating of 73 for a given test and answered 86% of the questions correctly their bias score would be −0.13 (0.73–0.86).

Within the path model analyses, average calibration was computed from all four tests. Change in calibration was computed by subtracting test one calibration from test four calibration.

Monitoring exercises

Each week students in the treatment class were given a monitoring worksheet (see Appendix B for sample) at the end of each class, with the exception of test dates and the introductory class (a total of 11 exercises). Each worksheet asked the students 1) to rate their understanding of the day's content on a 100-point scale, 2) what concepts from the class they found difficult to understand, 3) specifically, what they would do to improve their understanding of the concepts they identified as difficult, and finally 4) three multiple choice review questions of the day's material followed by a confidence judgment for each on a 100-point scale. GCJs were measured using students' ratings from the first question ("Please indicate below your overall understanding of the content from today's class) from the first ten monitoring worksheets. For the purpose of analysis composite scores were created by grouping the GCJs into 3 Units: Unit 1: weeks 1, 2, 3; Unit 2: weeks 4, 5, 6; Unit 3: weeks 7, 8, 9, 10. The review questions were answered and discussed before the class ended so that students could receive feedback from the teacher and compare their confidence estimate with their actual performance. The total time for each exercise was approximately five to ten min. The students then kept a copy of the monitoring

worksheet as part of a portfolio to be handed in at the end of semester. They were encouraged to regularly revisit the monitoring worksheets and use them to guide their study and review process along the semester. In addition to the weekly feedback on self-monitoring, the students were presented with monitoring feedback the week following each test in the form of their overall calibration and bias scores. Students were provided with Powerpoint slides and an accompanying verbal description of how to interpret these indexes after each test. This feedback provided an overall estimate of monitoring accuracy that students could use as information to assist in the regulation of future study habits and test performance. In contrast, the only form of intervention offered to the comparison group was the opportunity to self-generate feedback between their confidence ratings and performance outcomes after each of the four tests.

Procedures

During the first week of the course students in both the treatment and comparison classes were given the educational psychology pretest, the educational psychology self-efficacy inventory, and the Raven's test. Participants in both classes also completed the educational psychology self-efficacy inventory at the end of the course. The classes met separately once per week for 16 weeks.

We used a repeated-measures design to examine changes in calibration and performance between the treatment and comparison group. The first author taught both sections of the course which met on a weekly basis. Both the treatment and comparison classes made item-by-item confidence judgments on a 100-point scale during each of the four classroom tests. Students in both classes were allowed to examine the results of each test the following week. This included an opportunity to examine item-by-item performance along with their confidence ratings, and the opportunity to ask questions about any items they were unclear on. In addition, the treatment class completed the weekly monitoring exercises. Students in both conditions were informed that their monitoring accuracy was being measured with such activities. Every effort was made to have uniform experiences for both the treatment and comparison classes. Therefore, both classes received the same Powerpoint slides, classroom activities, and used the same textbook. Material was presented in parallel fashion across both classes. The course content contained a detailed unit on metacognition including monitoring that occurred during the first third of the course.

Results

Prior to performing analyses reported below, data were screened for outliers and normality. While performing the analyses, assumptions of the procedures were tested, and met unless specified. Additionally, during each analysis standardized residuals were examined to screen for multivariate and within-cell outliers, or data points with unusually strong influence. Except where noted, no multivariate or within-cell outliers were found. For repeated-measures comparisons between the treatment and comparison condition, performance and calibration scores were converted to standardized z scores to control for differential difficulty across tests, which has been shown to affect monitoring accuracy scores (Schraw & Roedel, 1994; Pressley & Ghatala, 1988). Higher z scores indicate higher test scores and more accurate calibration. In the following section results are reported according to the research questions described above.

Descriptive statistics

In Table 1 descriptive statistics (using raw scores) are presented for the major study variables. No significant differences were found between the treatment and comparison groups on the educational psychology pretest or the Raven's test for performance, calibration, or bias. In addition, there were no significant differences between the two groups in self-efficacy for learning educational psychology.[2]

Do distributed monitoring exercises and feedback improve monitoring accuracy?

Calibration of confidence judgments

In order to test the effect of the distributed training intervention on calibration a 2 (condition — experimental vs. comparison) by 5 (pretest through test 4) repeated measures analysis of covariance (RMANCOVA) was performed. Student scores on the Raven's Progressive Matrices test and student gender were covaried to account for individual differences in intellectual performance and gender differences in performance. One student with an extreme pretest score ($z = -3.82$) was removed from this analysis.

Results of the analysis showed significant differences wherein students in the experimental group scored significantly higher ($M = 0.19$) than the comparison group ($M = -0.52$, $F_{(1,74)} = 8.53$, $p < 0.005$, $\eta^2 = 0.10$). As expected there was a significant and strong interaction between condition and calibration ($F_{(4, 71)} = 3.61$, $p < 0.01$, $\eta^2 = 0.17$; no other effect was significant at $p < 0.05$). The details of this interaction are presented in Figure 1 showing that initial calibration ratings for both groups were equivalent, but as the experimental group received more monitoring exercises and feedback they quickly became more proficient at accurately calibrating their performance. Student performance in the experimental condition was consistently one standard deviation better than those in the comparison condition after the first test. Tukey HSD comparisons were performed comparing the two groups at each time point in order to determine which of the comparisons were significantly different. These follow-up analyses indicated that group differences in performance were significantly different at $p < 0.05$ for tests 2, 3, and 4, but not 1.

Do distributed monitoring exercises and feedback improve student outcomes?

Test scores

In order to test the effect of the distributed training intervention on test performance a 2 (condition — experimental vs. comparison) by 5 (pretest through

[2] The same instructor taught both classes. It is not the case that the instructor was blind to the conditions, and therefore one could argue that the results could be at least partly attributable to instructor bias, rather than the manipulated variable. Every precaution was made to equate student experiences in both conditions. The instructor used identical powerpoint presentations during lecture, both classes dealt with the same material at the same time, the same tests were administered to both classes, all tests and projects were scored using identical rubrics and answer keys, and so on. The instructor made a conscious effort to provide the same instructional experience to both classes. However, it is possible that bias is present. What that would mean is that instructor expectancies could influence the extent to which students self-monitor or the extent to which students become more adept at self-monitoring. That outcome itself is compelling and powerful, were that the case.

Table 1 Means and standard deviations of major study variables

Variable	Treatment group		Comparison group		Total	
	M	SD	*M*	SD	*M*	SD
Raven's score	0.84	0.17	0.78	0.16	0.82	0.17
Raven's calibration	0.28	0.13	0.30	0.11	0.29	0.12
Pretest score	0.51	0.10	0.48	0.09	0.50	0.10
Pretest calibration	0.46	0.05	0.46	0.06	0.46	0.05
Test 1 score	0.81	0.12	0.78	0.13	0.80	0.12
Test 1 calibration	0.30	0.10	0.30	0.10	0.30	0.10
Test 1 confidence	0.73	0.13	0.78	0.12	0.76	0.13
Test 1 bias	−0.08	0.15	0.00	0.15	−0.04	0.15
Test 2 score	0.91	0.09	0.82	0.11	0.86	0.11
Test 2 calibration	0.22	0.13	0.28	0.11	0.25	0.12
Test 2 confidence	0.82	0.13	0.80	0.12	0.81	0.12
Test 2 bias	−0.09	0.13	−0.01	0.16	−0.05	0.15
Test 3 score	0.85	0.09	0.70	0.15	0.78	0.14
Test 3 calibration	0.22	0.13	0.28	0.11	0.25	0.12
Test 3 confidence	0.76	0.16	0.69	0.19	0.73	0.18
Test 3 bias	−0.10	0.16	−0.01	0.20	−0.05	0.18
Test 4 score	0.84	0.08	0.80	0.10	0.82	0.09
Test 4 calibration	0.24	0.09	0.29	0.10	0.26	0.10
Test 4 confidence	0.83	0.12	0.79	0.14	0.81	0.13
Test 4 bias	−0.01	0.13	−0.01	0.15	−0.01	0.14
Schema score	0.88	0.05	0.83	0.08	0.86	0.07
Self-efficacy 1	32.69	3.90	32.26	4.38	32.49	4.11
Self-efficacy 2	29.47	4.81	30.15	5.37	29.79	5.06

Scores for test performance correspond to percent correct. Scores for confidence range from 0–100. Calibration scores can range from 0–1 with scores closer to zero being more accurate. Bias scores can range from −1 to +1 with negative scores representing underconfidence and positive scores representing overconfidence. Schema scores correspond to a grade ranging from 0 – 1.

test 4) repeated measures analysis of covariance (RMANCOVA) was performed. Student scores on the Raven's Progressive Matrices test and student gender were covaried to account for individual differences in intellectual performance and gender differences in performance.

Results of the analysis showed significant differences wherein students in the treatment group scored significantly higher ($M = 0.16$) than the comparison group ($M = -0.49$, $F_{(1,74)} = 12.15$, $p < 0.0001$, $\eta^2 = 0.16$). As expected there was a significant and strong interaction between condition and calibration ($F_{(4, 71)} = 4.21$, $p < 0.004$, $\eta^2 = 0.19$; no other effect was significant at $p < 0.05$). The details of this interaction are presented in Figure 2. The interaction in Figure 2 is similar to that of Figure 1, showing that initial test performance for both groups were equivalent, but as the treatment class received more practice and feedback they quickly improved their performance relative to the comparison class. Student performance in the experimental condition was consistently one standard deviation better than those in the comparison condition after the first test. Tukey HSD comparisons were performed comparing the two groups at each time point in order to determine which of the comparisons were significantly different. These follow-up analyses indicated that only performance on tests 2, 3, and 4 were significantly different at $p < 0.05$.

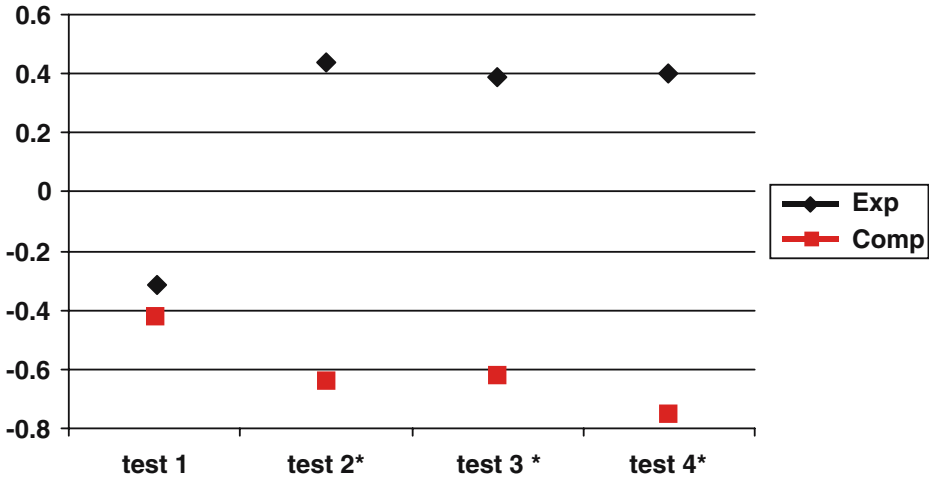

Figure 1 Change in calibration scores for the treatment and comparison class

Schema Representation scores

To test for significant differences in schema scores, a univariate ANCOVA was performed with condition as the independent variable, controlling for gender and Raven's scores. One student who was a within-cell outlier ($z = -4.08$) was removed. As expected, those students in the experimental condition ($M = 4.49$) scored significantly higher than those in the comparison condition ($M = 3.83$, $F_{(1,78)} = 8.55$, $p < 0.005$, $\eta^2 = 0.10$).

Does calibration or changes in calibration account for the observed changes in performance measures and self-efficacy over the semester?

The analyses for this section were guided by the theoretical path model presented in Figure 3. The goal of these analyses was to test the goodness of fit of this model to

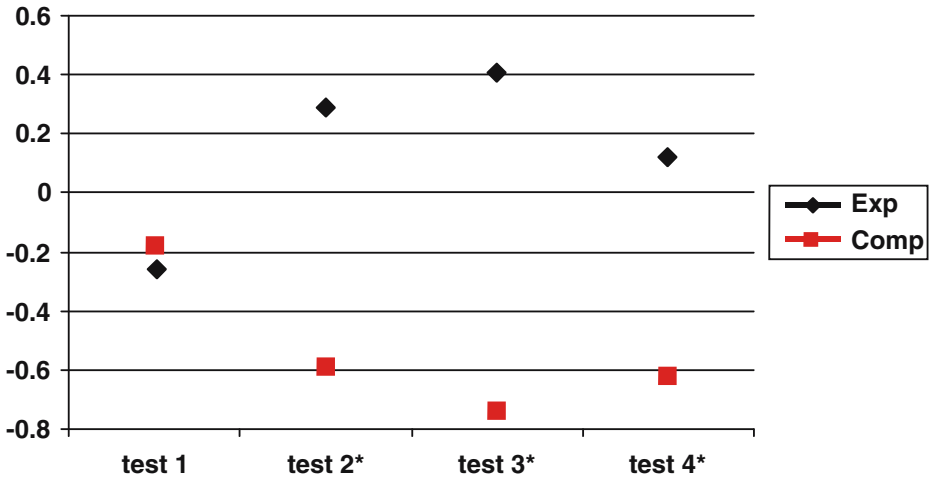

Figure 2 Change in test scores for the treatment and comparison class

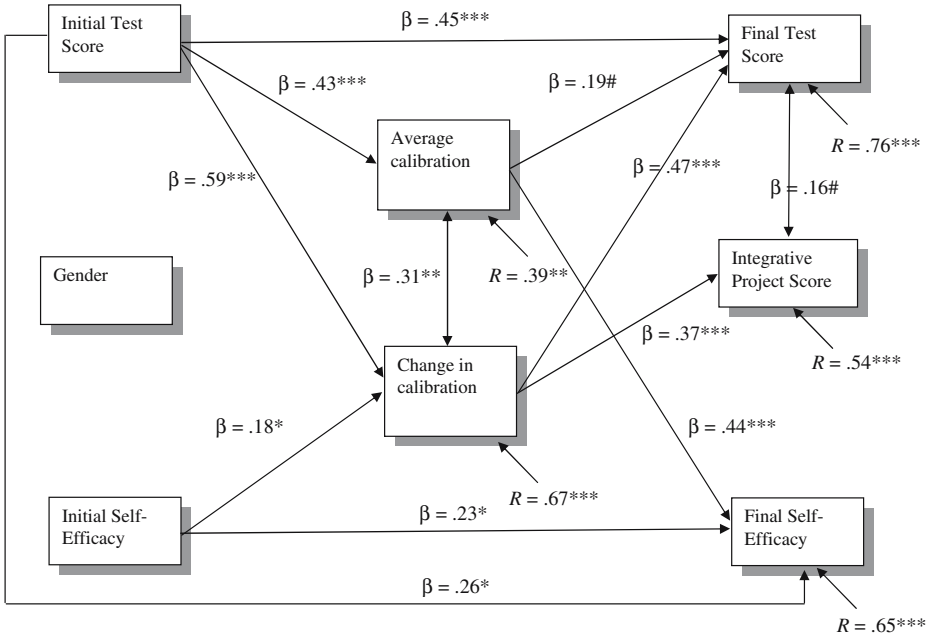

Figure 3 Path model including major study variables

explain how important course outcomes were influenced by calibration and changes in calibration that took place during the course for the treatment class.

In order to generate the path model presented in Figure 3, several multiple regression analyses were performed following the guidelines for causal modeling in Cohen and Cohen (1983). A path analysis approach was selected instead structural equation modeling due to the limited number of subjects. For each multiple regression assumptions were tested, and standardized residuals were checked for multivariate outliers (none were identified). All predictors were entered simultaneously.

In general, both calibration variables and all three outcomes were each used as dependent variables predicted from all antecedent or concomitant variables in the path model. Thus, the first multiple regression analysis predicted average calibration from (a) initial test score, (b) gender, (c) initial efficacy, and (d) change in calibration. When all variables were in the equation, betas and significance levels were reported on appropriate connecting lines. Another analysis was then performed predicting change in calibration from (a) initial test score, (b) gender, (c) initial efficacy, and (d) average calibration. Following this, three multiple regression analyses were performed predicting each of the three outcomes (final test score, integrative project score, and final efficacy) from all other antecedent variables (initial test score, gender, initial efficacy, average calibration, and change in calibration) and the other two outcome variables.

As Figure 3 shows, there are multiple direct and indirect effects. As predicted, calibration appears to at least partially account for academic outcomes (e.g., test score, integrative final project score), and also appears to partially account for the relationship between initial and final self-efficacy. It is important to note that these are unique effects that are substantial in magnitude once initial (baseline) variables are controlled for.

Specifically, while there is a significant direct effect of initial test score on the summative test score, there is also a significant direct effect of change in calibration, and an indirect relationship of initial test score through change in calibration. Thus, it appears that an increase in calibration has a substantial effect on summative test scores even once initial scores and other variables are taken into account.

Not surprisingly, similar effects were found when predicting the integrative project, with the exception of no unique direct effects from the exogenous variables. In this case, then, change in calibration has a unique direct effect upon student scores on the integrative project, while initial academic performance has only an indirect effect on this variable through calibration.

Finally, as with the previous two outcomes, there was both a direct effect of initial efficacy on final efficacy as well as a direct effect of average calibration. There was a very weak indirect effect of initial efficacy through change in calibration to final efficacy, but given the magnitude of the effect, it probably has minimal practical implications. Below, we report detailed results of questions generated by the path model.

Can calibration be responsible for academic performance?

One of the overarching questions in monitoring research is the extent to which we can attribute the level of academic performance or changes in the level of academic performance to monitoring ability. From the path model it is apparent that both average calibration and changes in calibration account for unique variance in student final test scores once initial test score and other variables were covaried. However, in order to test this question more specifically, we performed a multiple regression analysis predicting the change in test scores over the semester (i.e., change in relative standing over the semester; calculated as final score−initial score for each student) from both average calibration and change in calibration. Change in monitoring accuracy had a substantial relationship to this variable ($\beta = 0.61$, $p < 0.0001$), while average monitoring accuracy was not significantly related ($\beta = -0.07$, $p < 0.42$). Thus, students who improved their calibration also had a strong tendency to improve their performance on class tests.

Is calibration responsible for changes in self-efficacy?

It is also apparent from the path model that calibration might have some ability to explain changes in self-efficacy over the semester. Thus, as in the preceding analysis, we calculated changes in self-efficacy over the semester, and predicted that change from both average calibration and the change in calibration would account for unique variance in final self-efficacy. The analysis showed a significant effect for average accuracy monitoring ($\beta = 0.40$, $p < 0.0001$), but not change in monitoring accuracy ($\beta = 0.-21$, $p < 0.25$).

Cross-lag analyses

Temporal analyses were utilized in order to examine the alternative hypotheses that self-efficacy could cause calibration or performance. Thus, three multiple regressions were performed, predicting final self-efficacy, final calibration, and final test scores from initial self-efficacy, initial calibration, and initial test scores. The results were illuminating. In the first analysis all three baseline measures predicted final

Table 2 Correlations with pretest and post-test self-efficacy and performance and monitoring variables

	GCJ Unit 1	GCJ Unit 2	GCJ Unit 3	Test 1 score	Test 2 score	Test 3 score	Final exam score	Test 1 mon. acc.	Test 2 mon. acc.	Test 3 mon. acc.	Final exam mon. acc.
Ed. psych self-efficacy pretest	0.39**	0.06	0.02	0.11	−0.17	0.18	0.05	−0.17	−0.06	−0.17	−0.20
Ed. psych self-efficacy post-test	0.33*	0.48**	0.51**	0.48**	0.40**	0.26*	0.26*	−0.50**	−0.48**	−0.49**	−0.46**

self-efficacy (βs ranging from 0.25 to 0.28, all $p < 0.05$). However, for the second analysis (predicting final calibration), only initial calibration was a significant predictor ($\beta = 0.47$ $p < 0.001$). For the third analysis (predicting final test scores) only initial test scores was a significant predictor ($\beta = 0.46$, $p < 0.001$).

In addition to examining the relationship between self-efficacy and calibration on test scores within the path analysis the zero order correlations between initial and final self-efficacy with performance and monitoring variables were also examined. These findings revealed some interesting trends that are shown in Table 2 and Figure 4. The only significant correlation with initial self-efficacy was with the GCJ from Unit 1. On the other hand, final self-efficacy shows an increasing relationship with GCJs throughout the semester and consistent relationships with measures of performance and calibration on the classroom tests. These findings coupled with findings from the path analysis above indicate important shifts that appear to take place during the semester. Overall, these analyses revealed that students who increased their calibration, and hence their test scores, also tended to see at least modest increases in self-efficacy.

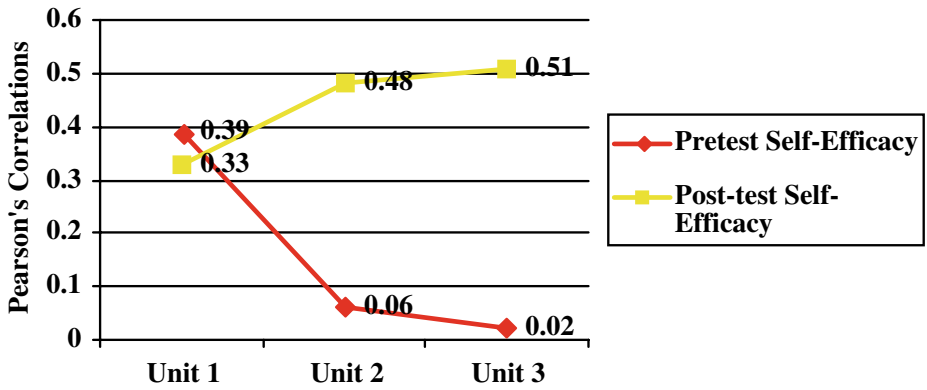

Figure 4 Correlations between JOLs and self-efficacy

Discussion

Results of this study revealed significant treatment effects of distributed monitoring exercises and feedback on calibration and performance. The treatment class differed from the comparison class with regard to calibration and performance at the level of one standard deviation. This difference emerged by the second of four tests and was maintained through the remainder of the course. This finding indicates that the monitoring exercises took time, in this case four weeks, to elicit their effects. Performance differences were also found between the treatment and comparison classes on the schema project, an integrative classroom project submitted at the end of the semester. These findings suggest that monitoring ability in the form of calibration impacts not only performance on multiple-choice tests but also on more authentic, performance-based measures targeting the integration of course content.

The group differences in calibration and test performance observed in this study demonstrate that the integration of metacognitive exercises within a curriculum in a distributed fashion was effective in promoting monitoring ability and academic achievement. This result extends findings by previous studies that have shown interventions to produce gains in calibration (Hacker et al., 2000; Nietfeld & Schraw, 2002) and positive effects of feedback in laboratory-based settings on calibration (West & Stanovich, 1997). It also yields support to the contention that metacognitive skills are trainable (Baker & Brown, 1984; Butler & Winne, 1995; Hartman, 2001; Koriat, 1997; Lin et al., 2002).

More importantly, our findings illustrate that distributed exercises were an appropriate approach in addressing the challenge of how to promote monitoring skills (Hacker et al., 2000; Magliano, Little, & Graesser, 1993; McCormick, 2003; Peverly & Brobst, 2003; Pintrich, Wolters, & Baxter, 2000; Sternberg, 2001; Underwood, 1961). The positive effect of this study confirms the similar results of explicitly teaching metacognitive skills to elementary school students (Desoete, Roeyers, & De Clercq, 2003). Our results also extend recent findings regarding the need to provide college students with external cues in order to ensure that they select effective study tactics (Winne & Jamieson-Noel, 2002).

The success of our approach at integrating metacognitive activities within the curriculum leads to important implications in promoting metacognitive skills in the classroom setting. The first is that the effect of the monitoring exercises was maintained throughout the semester despite an instructional focus on material that covered diverse topics. The positive effect of monitoring exercises on test performance in our study shows that promoting monitoring skills enhances mastery of the course content. One might expect similar or even more impressive effects in a well-structured subject area that allows continuous development of course content as opposed to one comprised of independent topics such as educational psychology. Second, the intervention produced large and consistent effects without taking a significant amount of time away from instruction. This time efficient intervention makes it easier for classroom teachers at all levels to adopt when they see that promotion of metacognition within their class can be achieved while they maintain instructional time simultaneously. The limitation of the design in the current study was that we were not able to isolate the effects of various components of the treatment condition (e.g., weekly monitoring exercises, calibration and bias feedback following each test).

Results of path analysis in this study indicate that calibration plays a significant role in not only objective measures of classroom performance (test performance)

but also on more authentic integrative knowledge (schema representation project) and on self-efficacy. These findings support current models of self-regulation that highlight the integral role of monitoring as an informant to processes that improve the understanding of new knowledge and impact the level of engagement in the activity (Butler & Winne, 1995; McCormick, 2003; Nelson & Narens, 1994; Schunk & Zimmerman, 2003; Schraw, 1998; Winne, 2001; Winne & Hadwin, 1998).

In this study neither the treatment nor the comparison class showed improvements in self-efficacy. Typically, educational psychology students show an increase in personal teaching efficacy (PTE) while advancing through their coursework (Hoy & Woolfolk, 1990; Nietfeld & Cao, 2003). We used a global measure of self-efficacy that was somewhat different from typical measures of PTE. We believe that educational psychology may be somewhat unique in that students enter the course with misconceptions about nature of the class. They tend to believe the course is easy and therefore might have inflated self-efficacy. Once immersed in the material, they realize the complexity of the topics and their self-efficacy might become more reasonable, hence decreased from the originally inflated levels.

In this study final self-efficacy showed an increasing relationship with weekly global confidence judgments throughout the semester and consistent relationships with measures of performance and calibration on classroom tests whereas initial self-efficacy only showed a significant relationship with GCJs from the first unit of the class. Given these findings, we would argue that as the course progresses self-efficacy is transformed into a belief that is informed more by monitoring processes and experienced-based beliefs (Koriat, 1997; Kelley & Jacoby, 1996). This finding supports the notion that self-efficacy is strongly influenced by domain-specific experience and successful navigation of challenges within the given domain (Bandura, 1997; Pajares, 1996). This argument is further supported by the relatively low but significant correlation between initial self-efficacy and final self-efficacy ($r = 0.33$). Thus, it is important to point out that although we did not find an improvement in self-efficacy at the group level, either between classes or with both classes aggregated, this does not diminish our findings on the key role that monitoring plays in final self-efficacy. Findings from both the path analysis and correlations with the GCJs converge to illustrate the interrelationship between monitoring and self-efficacy in the knowledge acquisition process during the course.

In sum, results of this study indicate that even modest interventions aimed at improving students' metacognitive monitoring can have significant payoffs in terms of calibration, performance, and self-efficacy. Moreover, our results suggest that calibration and changes in calibration during a course contribute a significant amount of variance in determining major course performance and affective outcomes.

The findings of this study provide leads for at least three areas of future research on monitoring. The first area for future investigations addresses the relationship between monitoring ability and knowledge domain. This study examined changes in performance and calibration for class content that consisted primarily of discrete units. We found an asymptotic effect in that the differences between the treatment and comparison groups appeared to level off. Investigations of monitoring within a well-structured subject area such as mathematics and chemistry, in which concepts and knowledge build one upon another continuously, will test whether intervention effects such as ours will continue over time.

A second area focuses on the complex relationship between metacognition and motivation. Since the ultimate goal of studying monitoring ability is to promote self-

regulated learning, which is mediated by both metacognitive and affective factors (Pintrich et al., 2000; Schunk, 1995; Schunk & Zimmerman, 2003), future investigations should address the interactions between monitoring ability and motivation. In particular, how does self-generated feedback, such as monitoring of comprehension and monitoring of performance, relate to students' goal setting and revision, assessment of academic progress, and adjustment of study strategies during the learning process.

A third area of future investigations is based on a developmental perspective that emphasizes the importance of developing metacognitive skills in the K-12 setting. Programs focusing on comprehension monitoring (Brown, Palincsar, & Armbruster, 1984; Brown, Pressley, Van Meter, & Schuder, 1996) have already been shown to be effective for elementary students in reading. Yet, to date there have been few attempts that aim at promoting K-12 level students' abilities to accurately estimate their academic work, assess their progress, and adjust their learning strategies. The distributed intervention approach applied in this study appears to be adaptable to a K-12 setting. Maintaining such interventions over extended periods of time may help younger students develop metacognitive abilities and enhance their performance.

Appendix A

Educational psychology self-efficacy questionnaire

This questionnaire is designed to help us gain a better understanding of the kinds of things that create difficulties for students in the educational psychology course. Please indicate your opinion about each of the statements below in reference to your current situation. Your answers are confidential and will not be identified by name.

The following statement is _____ like me.

1 = Nothing, 2 = very little, 3 = somewhat, 4 = quite a bit, 5 = a great deal

___1. I am sure that I can learn educational psychology.
___2. I can get a good grade in educational psychology.
___3. I am sure I could do advanced work in educational psychology.
___4. I have a lot of self-confidence when it comes to educational psychology.
___5. I am not the type to do well in educational psychology.
___6. It takes me a long time to catch on to new topics in educational psychology.
___7. Even before I begin a new topic in educational psychology, I feel confident I'll be able to understand it.
___8. I think I have good skills and strategies to learn educational psychology.

Appendix B

Weekly monitoring exercise sheet

Research/cognitive development

Please indicate below your overall understanding of the content from today's class:
0% Accurate——————————————————100% Accurate

- What concept(s) from today's class did you find difficult to understand?
- Specifically, what will you do to improve your understanding of the concept(s) you listed above?

1. Experimental research requires which one of the following?

   A. Manipulating an aspect of the environment
   B. Being able to predict two or more variables
   C. Describing each variable in considerable detail
   D. Studying behavior in an actual classroom environment

0% Accurate——————————————————100% Accurate

2. Mr. Johnson teaches a class of 20 8-year-old third graders. His goal for the upcoming school year is to help at least 50% of his students reach formal operations. Judging from Piaget's theory, we would expect that Mr. Johnson's goal is:

   A. An easy one to attain
   B. Almost impossible to attain
   C. Attainable only if he emphasizes abstract reasoning throughout the school year
   D. Attainable only if his students have had enriched educational experiences most of their lives

0% Accurate——————————————————100% Accurate

3. From a Vygotskian perspective, scaffolding serves what purpose in instruction?

   A. It gives an idea of what they need to do to get good grades
   B. It keeps school tasks within their actual developmental level
   C. It lets them learn by watching one another
   D. It supports them as they perform difficult tasks

0% Accurate——————————————————100% Accurate

References

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

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: W.H. Freeman.

Bandura, A., & Cervone, D. (1986). Differential engagement of self-reactive influences in cognitive motivation. *Organizational Behavior and Human Decision Processes, 38*, 92–133.

Brown, A. L., Palincsar, A. S., & Armbruster, B. B. (1984). Instructing comprehension-fostering activities in interactive learning situations. In H. Mandl, N. L. Stein, & T. Trabasso (Eds.), *Learning and comprehension of text* (pp. 255–286). Hillsdale, New Jersey: Erlbaum.

Brown, R., Pressley, M., Van Meter, P., & Schuder, T. (1996). A quasi-experimental validation of transactional strategies instruction with previously low-achieving, second-grade readers. *Journal of Educational Psychology, 88*, 18–37.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245–281.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*, 404–431.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, New Jersey: Erlbaum.
Delclos, V. R., & Harrington, C. (1991). Effects of strategy monitoring and proactive instruction on children's problem-solving performance. *Journal of Educational Psychology, 83*, 35–42.
Desoete, A., Roeyers, H., & De Clercq, A. (2003). Can offline metacognition enhance mathematical problem solving? *Journal of Educational Psychology, 95*, 188–200.
Dunlosky, J. (2004). Metacognition. In R. R. Hunt & H. C. Ellis (Eds.), *Fundamentals of cognitive psychology* (7th edn.). New York: McGraw-Hill College.
Ghatala, E. S., Levin, J. L., Pressley, M., & Goodwin, D. (1986). A componential analysis of the effects of derived and supplied strategy-utility information on children's strategy selection. *Journal of Experimental Child Psychology, 41*, 76–92.
Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv–xxviii). Hillsdale, New Jersey: Earlbaum.
Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 702–718.
Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*, 160–170.
Hartman, H. J. (2001). Developing students' metacognitive knowledge and skills. In H. J. Hartman (Ed.), *Metacognition in learning and instruction* (pp. 33–67). Dordrecht, Netherlands: Kluwer.
Hoy, W. K., & Woolfok, A. E. (1990). Socialization of student teachers. *American Educational Research Journal, 27*, 279–300.
Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35*, 157–175.
Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Pscyhologica, 77*, 217–273.
Koriat, A. (1997). Monitoring one's knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370.
Lin, L. M., Zabrucky, K. M., & Moore, D. (2002). Effects of text difficulty and adults' age on relative calibration of comprehension. *The American Journal of Psychology, 115*, 187–198.
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., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*, 723–731.
McCormick, C. B. (2003). Metacognition and learning. In W. M. Reynolds & G. E. Miller (Eds.), *Handbook of psychology: Educational psychology* (pp. 79–102). Wiley.
Metcalfe, J. (2000). Metamemory: Theory and data. In E. Tulving & F. I. M. Craik (Eds.), *Oxford handbook of memory* (pp. 197–211). London: Oxford University Press.
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: Comments on Schraw (1995). *Applied Cognitive Psychology, 10*, 257–260.
Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–27). Cambridge, Massachusetts: MIT.
Nietfeld, J. L. (2002). Beyond concept maps: Using schema representations to assess pre-service teacher understanding of effective instruction. *The Professional Educator, 25*(1), 15–27.
Nietfeld, J. L., & Cao, L. (2003). Examining instructional strategies that promote pre-service teachers' personal teaching efficacy. *Current Issues in Education* [On-line], *6*(11). Available: http://cie.ed.asu.edu/volume6/number11/.
Nietfeld, J. L., Enders, C. K., & Schraw, G. (2006). A Monte Carlo comparison of measures of relative and absolute monitoring accuracy. *Educational and Psychological Measurement, 66*, 258–271.
Nietfeld, J., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research, 95*, 131–142.
Osborne, J. (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation, 8*(6). Available online: http://ericae.net/pare/getvn.asp?v=8&n=6.
Pajares, F. (1996). Self-Efficacy beliefs in academic settings. *Review of Educational Research, 66*, 543–578.
Pajares, F., & Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in mathematical problem-solving. *Contemporary Educational Psychology, 20*, 426–443.
Paris, S. G., Cross, D. R., & Lipson, M. Y. (1984). Informed strategies for learning: A program to improve children's reading awareness and comprehension. *Journal of Educational Psychology, 76*, 1239–1252.

Peverly, S. T., & Brobst, K. E. (2003). College adults are not good at self-regulation: A study on the relationship of self-regulation, note taking, and test taking. *Journal of Educational Psychology, 95*, 335–346.
Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33 – 40.
Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). Assessing metacognition and self-regulated learning. In G. Schraw & J. C. Impara (Eds.). *Issues in the measurement of metacognition* (pp. 43–97). Lincoln, Nebraska: Buros Institute of Mental Measurements.
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. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist, 25*, 19 – 33.
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.
Raven, J. C. (1962). *Advanced progressive matrices, set II*. London: H.K. Lewis (Distributed in the United States by the Psychological Corporation, San Antonio, Texas).
Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*. New York: Academic.
Schneider, W., & Pressley, M. (1997). *Memory development between two and twenty* (2nd edn.). Hillsdale, New Jersey: Erlbaum.
Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology, 19*, 143–154.
Schraw, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem. *Applied Cognitive Psychology, 9,* 321–322.
Schraw, G. (1998). On the development of adult metacognition. In C. M. Smith & T. Pourchot (Eds.), *Adult learning and development: Perspectives from educational psychology* (pp. 89–106). Mahway, New Jersey: Erlbaum.
Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*, 351–371.
Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology, 18*, 455 – 463.
Schraw, G., & Roedel, T. D. (1994). Test difficulty and judgment bias. *Memory & Cognition, 22*, 63 – 69.
Schunk, D. H. (1995). Self-efficacy and education and instruction. In J. E. Maddux (Ed.), *Self-efficacy, adaptation, and adjustment: Theory, research, and application* (pp. 281–303). New York: Plenum.
Schunk, D. H., & Zimmerman, B. J. (2003). Self-regulation and learning. In W. M. Reynolds, G. E. Miller, & I. B. Weiner (Eds.), *Handbook of psychology* (vol.7). Wiley.
Sternberg, R. J. (2001). Metacognition, abilities, and developing expertise: What makes an expert student? In H. J. Hartman (Ed.), *Metacognition in learning and instruction* (pp. 229–260). Netherlands: Kluwer.
Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66 – 73.
Underwood, B. J. (1961). Ten years of massed practice on distributed practice. *Psychological Review, 68,* 229–247.
West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin & Review, 4*, 387–392.
Winne, P. H. (2001). Information processing models of self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theory, research, and practice*. New York: Longman.
Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 227–304). Mahwah, NJ: Erlbaum.
Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology, 27*, 551–572.
Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, New Jersey: Prentice-Hall.