Hacker, D. J., Bol, L., & Keener, M. C. (in press). Metacognition in education: A focus on calibration. In J. Dunlosky, & R. Bjork (Eds.), *Handbook of Memory and Metacognition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Running Head: METACOGNITION IN EDUCATION

Metacognition in Education: A Focus on Calibration

Douglas J. Hacker

University of Utah

Linda Bol

Old Dominion University

Matt C. Keener

University of Utah

Metacognition in Education: A Focus on Calibration

"Why investigate metacognition?" Thomas Nelson and Louis Narens asked this question in the title to a chapter they authored in 1994. Their question was not asked in a disparaging way, but was intended to encourage reflection on the reasons for the lack of "cumulative" progress in research on learning and memory over the last half century. Nelson and Narens speculated that this lack of cumulative progress was due, in part, to three shortcomings: (a) lack of a target for research, (b) overemphasis on a nonreflective-organism approach, and (c) short-circuiting via experimental control (i.e., researchers' attempts to control variations in participants' self-directed cognitive processing). Of these three shortcomings, the first was the inspiration for the present chapter.

Nelson and Narens explained that a target for research "should be defined in terms of some to-be-explained behavior of a specific category of organism in a specific kind of environmental situation" (p. 3). In their own work, they addressed this "lack of a target for research" by specifically identifying the to-be-explained behavior as mnemonic behavior, the specific category of organism as college students, and the environmental situation as studying for and taking an examination. They went on to argue that targets for research in the area of learning and memory typically have been restricted to the laboratory, and that although there is a continued need for laboratory work, there is also a need for researchers to go outside the laboratory into more ecologically valid environmental situations. A quotation from the Nelson and Narens chapter, provided by Parducci and Sarris (1984, pp. 10-11), aptly encapsulates this view, "The desire for ecological validity…. cannot be separated from the concern to make psychology more practical…. Scientists continue to study psychological problems without apparent concern for practical applications…. There do seem to be strong forces pushing even

traditional areas of psychological research in practical directions." We have resonated strongly with Nelson and Narens's arguments, and in this chapter, we have followed their guidelines in identifying a target for research: Calibration is the to-be-explained behavior; students—elementary to graduate—constitute the specific category of organism; and the classroom is the environmental situation.

Our plan for this chapter is first to expand on Nelson and Narens's argument to go outside the laboratory into more naturalistic environmental situations to study learning and memory. The environment in which metacognition is examined can impact the results of studies and therefore can impact our notions of the general character of metacognition. Second, we will present a brief overview of Nelson and Narens's (1990) model of metacognition for the purpose of describing the metacognitive monitoring and control processes that potentially interact in educational contexts. Last, as mentioned previously, we will narrow our focus on metacognition in education to calibration, how it is measured, and the calibration of students in classroom contexts.

Laboratories versus Classrooms

A common practice of researchers who conduct laboratory studies in learning and memory is to generalize their results to educational contexts. Discussion sections often provide suggested educational implications, some of which may be readily and productively applied to educational contexts, others that are not likely practical, and still others that are intended only as a call for future research. We are not advocating that learning and memory researchers should stop this practice. Providing educational implications should be a major concern for psychologists wishing to make their work more applicable to "naturalistic contexts" and can be quite helpful to researchers and practitioners interested in improving learning environments.

However, generalizing findings from studies that have used content and procedures that have little resemblance to actual classroom practices is risky and in some cases may be unwarranted (Lundeberg & Fox, 1991; McCormick. 2003; Winne, 2004). In a laboratory context, the goal is to control materials, procedures, participants, and experimental conditions, and the greater extent to which control can be achieved the more certain researchers can be that causes for thought or behavior have been identified. In the area of metacognition, this experimental rigor has been applied to a limited range of learning, most often including feeling of knowing (FOK), ease of learning (EOL), judgments of learning (JOL), confidence in retrieval, allocation of study time, or comprehension of short narrative or expository texts (Nelson & Narens, 1994).

In naturalistic contexts, especially classroom contexts, such controls are difficult to manage. Conditions for learning are massively complex in comparison to laboratories. Information can be encoded in multiple ways, including but not limited to lecture, reading, participation in group discussions, question and answer, and in some cases by physically manipulating materials (Maki & McGuire, 2002). Moreover, in general, students are likely more motivated to perform well on a classroom test that is going to contribute to their overall grade for a course than for a test that has little long-term consequence for them. And, the interval between learning and testing in a classroom context can be considerably longer than in a laboratory context in which it is often the case that barely an hour passes between learning and testing. In sum, differences between laboratory and classroom contexts entail not only the type of learning but the depth, breadth, and motivation for learning, all of which can impact one's ability to monitor and control learning.

Space does not permit an extensive analysis of the issues surrounding generalizability between laboratory and classroom contexts, or between different classroom contexts. However, allow us to provide an illustration that may shed additional light on some of the issues.

Lundeberg and Fox (1991) conducted a meta-analysis of laboratory and classroom studies investigating a form of metacognition called the *test expectancy effect*. The test expectancy effect was first reported by Meyer (1934) who found that students who were expecting to receive an essay test performed better on both an essay test and a multiple-choice test than students who were expecting to receive a multiple-choice test. Since then, the recommended study skill strategy has been to prepare for an essay test regardless of the actual type of test a person is to receive. Lundeberg and Fox's (1991) results showed that the test expectancy effect was true, but only for studies that were conducted in laboratory contexts. In studies conducted in classroom contexts, the exact opposite result was found. As a result of their meta-analysis, Lundeberg and Fox recommended that "In the classroom, the simplest advice, akin to the encoding specificity view, would be: Study for the type of test you expect to receive" (p. 97).

In addition to the practical advice that can be garnered from this study, the results point directly to our argument that generalizing findings from laboratory studies of metacognition to classroom contexts can at times be risky. Before such generalization can occur, there needs to be a better understanding of the factors that contribute to metacognitive judgments concerning the selection and use of study strategies and the conditions under which those judgments are made. If the conditions in a laboratory context approximate conditions in classrooms, generalizing from from one to the other would not be controversial. However, if conditions differ, and they likely do, factors that are known to affect metacognitive judgments in classroom contexts (e.g., depth

and breadth of knowledge, input from co-learners, motivation, or the social comparisons that learners make in a social setting) will need to be introduced and controlled in the laboratory. Until these factors are more thoroughly investigated, one should be cautious about generalizing from the laboratory to the classroom.

Metacognitive Monitoring and Control

Nelson and Narens (1990) proposed a theoretical framework for metacognition that has served well as a description of the components and processes that comprise this concept. Their framework is based on three principles: (a) Mental processes are split into an object-level (i.e. cognition) and a meta-level (i.e., metacognition), (b) the meta-level contains a dynamic model of the object-level, which is the source of metacognitive knowledge or understanding of the object-level , and (c) there are two processes corresponding to the flow of information from the object-level to the meta-level (i.e., monitoring) and from the meta-level to the object-level (i.e., control). Metacognition can be viewed as monitoring and control of a lower level of thought by a higher level of thought (Broadbent, 1977). Through monitoring, people obtain information at the metacognitive level about the status of knowledge or strategies at a cognitive level; and through control, people can use their metacognitive knowledge or understanding at the metacognitive level to regulate thought at the cognitive level (Hacker, 1998, 2004).

To illustrate the dynamic interplay between monitoring and control, consider calibration. In brief, calibration is a measure of the degree to which a person's judged ratings of performance correspond to his or her actual performance (Keren, 1991; Lin & Zabrucky, 1998; Winne, 2004; Yates, 1990). Although there are several significant contributors to calibration accuracy, the underlying psychological process reflected in calibration entails a person's monitoring of what he or she knows about a specified topic or skill and judging the extent of that knowledge in

comparison to some criterion task, such as an examination. For instance, while studying for an hour or two for an upcoming chemistry test on chemical nomenclature, students may continuously monitor what they know and judge that more studying is necessary to get a decent grade. They can exert further control over their studying for several more hours at which time they will again monitor what they know and judge that a grade of about 90% correct is possible and acceptable. That judgment of 90% is then compared to their actual performance, which for illustrative purposes turns out to be 95% correct. Calibration in this case is the difference between the judged 90% and the actual 95% correct, which indicates not only that the students were fairly accurate in monitoring their knowledge but that they were slightly underconfident.

This example illustrates how people, as agents of their own thoughts and behaviors, can monitor their knowledge or skills, establish their own goals for learning, develop plans to achieve their goals, control the deployment of those plans, monitor the progress of their plans, further control the plans if necessary, and judge when they have been achieved. In other words, people can be self-regulators of their own behaviors (Zimmerman, 2000). Thus, this example also highlights the importance of calibration in educational contexts. As a further illustration, consider how inaccurate calibration during reading could sway students to ineffectively regulate their learning of text (Lin & Zabrucky, 1998). On the one hand, strong overconfidence during reading could fail to trigger appropriate control processes necessary for students to attain greater comprehension of the text. On the other hand, strong underconfidence could cause students to misallocate precious study time to continue reading in the hopes of further comprehending the text when in fact their comprehension may be more than sufficient for the task.

In summary, the Nelson and Narens's (1990) theoretical framework of metacognition provides important insights into the dynamic interplay that exists between monitoring and

control processes as people attempt to influence their learning and memory. Although this theoretical framework is based almost entirely on laboratory research, the classroom context provides fertile ground for the application of theory to practice. At a minimum, to become self-regulated learners, students at the metacognitive level need to accurately monitor their ongoing cognitive states and processes, and the information obtained from such monitoring must be used to exert control to regulate those cognitive states and processes. The importance of accurate monitoring and control in relation to calibration has been succinctly summarized by Winne (2004), "Learning will be inversely proportional to the degree of calibration bias and proportional to calibration accuracy" (p. 476).

A Focus on Calibration

At this point, we would like to focus our attention more squarely on calibration, which is a type of metacognition that has been investigated perhaps more extensively in educational contexts than other types of metacognition. In the sections that follow, we intend to give a fuller description of calibration, describe the various ways in which it is measured, more fully discuss the importance of calibration to learning and memory in educational contexts, and describe patterns of findings in classroom contexts. We will end with a discussion of directions for future research.

What is Calibration

Calibration is the degree to which a person's perception of performance corresponds with his or her actual performance (Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1982; Nietfeld, Cao, & Osborne, 2006). In other words, learners make judgments about what knowledge or skill they have learned, and those judgments are compared to an objectively determined measure of that knowledge or skill (Winne, 2004; for other measures of judgment accuracy, please see

Benjamin & Diaz, this volume). As in the example given earlier, a student can monitor his or her learning before testing and make a prediction that 90% of the to-be-tested material has been mastered. In addition, the student's subjective judgment concerning what material has been mastered can occur after testing. Monitoring judgments that follow performance are commonly called postdictions (Lin & Zabrucky, 1998).

Nelson and Narens (1994) drew a distinction between prospective monitoring judgments and retrospective monitoring judgments that clarifies the distinction between prediction judgments and postdiction judgments. Figure 1 (adapted from Nelson & Narens, 1994) shows three stages of learning (i.e., acquisition, retention, and retrieval), the various monitoring judgments that a person can make (e.g., judgments-of-learning, feeling of knowing), and the control processes that are informed by monitoring (e.g., allocation of study time, termination of study). We have added to this figure where we believe prediction and postdiction judgments fit within the stages of learning. A prediction judgment is a monitoring judgment that comes after acquisition and retention but prior to retrieval; a postdiction judgment follows retrieval. Therefore, predictions can be thought of as prospective monitoring judgments (i.e., a person monitors his or her knowledge or skill before retrieval of the knowledge or skill). In some respects, a prediction judgment is a type of self-efficacy judgment (Hertzog, Dixon, & Hultsch, 1990) in that the magnitude of the judgment reflects a person's belief in his or her mastery of some learning or memory task. A postdiction judgment can be thought of as a retrospective monitoring judgment (i.e., a person monitors his or her knowledge or skill after retrieval). Both judgments can be used to inform control processes (Nelson & Narens, 1990, 1994). Optimistic predictions may lead people directly into retrieval, believing they have mastered the material or skill; pessimistic predictions may convince people they need to return to acquisition and

retention. Postdictions, which overlap to some degree with "confidence in retrieved answers," provide learners with more accurate feedback on their monitoring proficiency (Maki, 1998; McCormick, 2003; Pressley & Ghatala, 1990). Based on this feedback, learners may employ different control processes during their next acquisition and retention task.

Insert Figure 1 about here

An important distinction must be made between calibration, which is referred to as absolute accuracy, and resolution or discrimination, which are referred to as relative accuracy. The two types of accuracy are often confused, although they represent two very different aspects of metacognitive monitoring and are measured in very different ways (Nelson, 1996). In a recent study by Maki, Shields, Wheeler, & Zacchilli (2005), in which absolute and relative accuracy were compared, no significant correlation was found between the two, suggesting that the two types of accuracy tap different metacognitive processes.

Absolute accuracy (aka calibration) refers to the degree of correspondence between a person's judged level of performance and his or her actual performance. Calibration judgments provide important estimates of overall memory retrieval; however, they do not provide good discrimination between what a person may or may not know. Relative accuracy does this by providing a measure of the degree to which a person's judgments can predict the likelihood of correct performance of one item relative to another (Nelson, 1984, 1996) or whether a target event will or will not occur (Yates, 1990). In other words, relative accuracy provides a measure of whether a person can discriminate between what is known or not known, whereas absolute accuracy indicates whether a person can estimate actual overall test performance (Nelson, 1996; Nietfeld, Cao, & Osborne, 2005; Nietfeld, Enders, & Schraw, 2006).

In educational contexts, measures of absolute accuracy tend to show better reliability than measures of relative accuracy and are more likely to show stable individual differences (Maki et al., 2005). Nevertheless, both measures can be quite useful. Calibration provides important estimates of overall recall levels, and relative accuracy provides important estimates of which items are correct or incorrect. Maki and colleagues argue that if students are overconfident in their predicted performance, they may prematurely end studying, thinking that they have mastered the to-be-tested material. Moreover, those same students may not know which specific topics within the to-be-tested material need further study. Thus, inability to predict overall test performance and to discriminate among known and unknown topics can have dire consequences for achievement on tests.

How Calibration is Measured

Although there is one commonly used measure of relative accuracy, that is, gamma (Maki et. al., 2005; Nelson, 1984, 1996; Wright, 1996), there are a variety of methods used to measure calibration. These methods can be grouped roughly into two categories: difference scores and calibration curves. Difference scores all involve taking the difference between judged performance and actual performance; however, there are at least four questions that should be considered: (a) What kind of judgment is being made, (b) what level of performance is being judged, (c) when is the judgment being made, and (d) how is the difference between judged and actual performance calculated? First, judgments can be made on a percentage of likelihood scale or confidence scale (i.e., 0%--no likelihood or confidence in knowing, 20% chance or confidence in knowing up to 100% chance or confidence in knowing). Often participants are restricted to 6 probabilities (0, 20, 40, 60, 80, 100), but in other cases are given a choice to select any value along a continuous line, with 0% at one end and 100% at the other (Schraw, Potenza, &

Nebelsick-Gullet, 1993). Judgments also can entail asking participants to state how many items they expect to get correct out of the total number of items (e.g., Of the 35 items, how many do you expect to get correct). Second, judgments can be directed at a local level (e.g., the mean of the judgments made on individual items on a test) or at a global level (i.e., all the items as a whole) on a test (Schraw, 1994). Third, as discussed earlier, judgments can be made before or after performance, that is, predictions or postdictions, respectively (Pressley & Ghatala, 1989; Pressley, Levin, Ghatala, & Ahmad, 1987; Pressley, Snyder, Levin, Murray, & Ghatala; See also Lin & Zabrucky, 1998 for predictions and postdictions made in calibration of comprehension studies).

Finally, the difference between judged and actual performance can be calculated in several ways. Perhaps the most straightforward measure of calibration concerns global-level judgments in which the absolute value of the difference between judged and actual performance is calculated (e.g., Hacker, Bol, Horgan, & Rakow, 2000; Pressley & Ghatala, 1989; Pressley, Levin, Ghatala, & Ahmad, 1987; Pressley, Snyder, Levin, Murray, & Ghatala, 1987). For instance, students will be asked to predict or postdict their performance by making a judgment on how many items on a test they expect to get correct or got correct, respectively. Once their actual performance is assessed, their actual scores are subtracted from their predicted and postdicted judgments, and the absolute value of that difference is taken. Values closer to zero indicate greater accuracy. If the absolute value is not taken, the resultant differences produce a bias score. That is, negative values indicate underconfident judgments and positive values indicate overconfident judgments. A student who predicts a score of 80 but actually scores a 70 would be overconfident and positively biased.

Measures of calibration involving local-level judgments are a bit more complicated but still relatively straightforward (See Keren, 1991 or Yates, 1990 for a detailed description of these measures). For each item, participants are asked to predict or postdict their performance. These predictions or postdictions are usually given as a confidence judgment expressed as a probability statement in answering the item correctly (e.g., 75% confident that I will get the answer correct). Performance is assessed with a "0" being assigned to incorrect items and a "1" to correct items. Calibration is calculated by taking the absolute value of the difference between the confidence judgment (expressed as a proportion) and performance. The differences calculated for the individual items are then summed, and this sum is divided by the total number of items. People are said to be well calibrated if in the long run, their assigned probabilities to the items are equal to their performance on the items (Lichtenstein, Fischhoff, & Phillips, 1982). Thus, the closer to zero the mean difference score is, the better calibrated a person is. A bias score also can be calculated at the item level by calculating the mean probability judgment and subtracting from it the mean performance score. Negative values indicate overall underconfidence and positive values overconfidence. Yates (1990) also suggests squaring the differences between probabilities assigned to each item and actual scores, producing a probability score (aka quadratic score or the Brier score). The mean probability score then can be used to assess calibration accuracy (See Yates, 1990 for a discussion of standards of accuracy).

The other method for measuring calibration is the calibration curve or graph (Keren, 1991; Yates, 1990). Actual performance is plotted on the y-axis, and predicted or postdicted performance is plotted on the x-axis. The 45 degree line represents perfect calibration in which predictions or postdictions are exactly equal to actual performance. Points below perfect accuracy indicate overconfidence, and points above indicate underconfidence. Calibration

graphs provide easily interpretable representations of the ways in which accuracy varies across performance levels rather than a single measure of the relation between predictions or postdictions and actual performance (Weingardt, Leonesio, & Loftus, 1994). Moreover, calibration graphs demonstrate the ways in which overconfidence and underconfidence in judgments vary with performance.

Figure 2 is a calibration graph that reflects calibration of test performance in a classroom context (Hacker, Bol, Horgan, & Rakow, 2000). In this case, the values on the y-axis represent students' actual proportion correct on the first of three tests, and the values on the x-axis represent students' predicted and postdicted scores on the test expressed as proportions. The five groups are approximate groupings representing students' overall academic performance across the semester-length course, with Group One earning As, Group Two Bs, Group Three Cs, Group Four Ds, and Group Five Fs. Predicted scores are represented by the hollow squares, and postdicted scores are represented by the filled circles.

As can be seen in this figure, as a general rule, predictions tend to exceed postdictions, and postdictions tend to be more accurate than predictions, although in this example the highest performing group is an exception. Greater accuracy of postdictions over predictions is a common finding in calibration research, and Pressley and Ghatala (1990) referred to this phenomenon as the *testing effect.* What is striking about the results displayed in the figure is that higher-achieving students tend to be underconfident in their predictions and postdictions, whereas lower-achieving students tend to be overconfident, with their predictions grossly overconfident.

Insert Figure 2 about here

In sum, all measures of calibration provide a quantitative assessment of the degree of discrepancy between perceived performance and actual performance. The discrepancy can be calculated at the item level and averaged over multiple items, or the discrepancy can be calculated at global levels in which students are asked to make a single judgment over multiple items. The closer to zero the discrepancies become, the better calibrated a person is said to be, with perfect calibration attained when the discrepancies are zero. A person is overconfident if the calculated discrepancies are positive values and underconfident if they are negative. In educational contexts, the general finding observed has been that underconfidence is associated with higher performance and overconfidence with lower performance.

Why is Calibration Important in Educational Contexts?

In many professions, the inability to make accurate, realistic predictions can have dire consequences (Allwood & Granhag, 1999; Dunning, Heath, & Suls, 2004). Such dire consequences are exemplified by a physician who is unrealistically confident in her diagnoses, a lawyer who may be unduly optimistic when predicting the verdicts of his court cases, or an airline pilot who overestimates her ability to handle challenging weather conditions. In classrooms, although the consequences of overconfidence or underconfidence may not be life threatening, they may certainly affect students' academic achievement and motivation. Students who are strongly underconfident may fail to disengage from studying for a test and misallocate precious study time because they assume that they have not mastered the material (Maki et al., 2005). Strong overconfidence while employing a specific learning strategy can provide a false sense of the strategy's effectiveness (Hacker, 1998). And relatedly, students could intentionally inflate their overconfidence during test preparation as a self-handicapping strategy that provides

a ready excuse when performance is poor (Winne, 2004). For example, "I studied really hard for the test, so the teacher must have given an unreasonably difficult test."

In an era of high-stakes accountability, the ability to perform well on tests has become increasingly important (Bol & Nunnery, 2004). Student performance on high-stakes tests impacts educational placements, grade promotion, academic major, college admissions, graduation, and entry into various professions. Therefore, students' ability to judge how well they have studied for an exam and how well they are likely to perform on the exam, as well as how well they can monitor performance during the exam, are essential skills contributing to their performance. Inaccurate calibration judgments have been linked to poor performance on various types of exams (e.g., Barnett & Hixon, 1997; Bol & Hacker, 2001; Bol, Hacker, O'Shea, & Allen, 2005; Hacker et al., 2000; Kruger & Dunning, 1999; Nietfeld et al., 2005). Thus, there is good evidence suggesting that if students are unable to produce accurate calibration judgments, they may not take the remedial steps necessary to promote their achievement or carefully evaluate their responses during or after the exam.

Overconfidence in judging one's knowledge, skill, comprehension, or test preparedness is a robust phenomena observed across many subject areas (e.g., Allwood & Granhag, 1999; Bol & Hacker, 2001; Bol et al., 2005; Dunning, Heath, & Suls, 2004; Flannelly, 2001; Glenberg, Wilkinson, & Epstein, 1982; Grimes, 2002; Hacker et al., 2000; McCormick, 2003; Nelson, 1999). To further complicate matters, Winne and Jamison-Noel (2002) have shown that students also can be biased with respect to their self-reporting of study techniques. They found that students appeared overconfident in their self-reports of whether their studying was guided by objectives and a planned method of studying. Thus, drawbacks associated with overconfident

predictions may be compounded by overconfident self-appraisals regarding the efficacy of any particular study strategy employed.

Overconfidence may influence attention or preparation more selectively. Students may not allocate their study efforts to those topics for which they are least prepared. Many studies have shown that students tend to be more overconfident when the material or test items are difficult and underconfident when the material or test items are easy, a phenomenon dubbed the "hard-easy" effect (Flannelly, 2001; Juslin, Winman, & Olsson, 2000; Nietfeld, et al., 2005; Winne, 2004; Winne & Jamison-Noel, 2002). Therefore, students may allocate the least amount of time to difficult material that is, ironically, most in need of additional study effort due to their unrealistic confidence judgments. In testing situations, students may not critically reconsider their responses because they are unjustifiably confident in their knowledge (Flannelly, 2001). Because students need to feel a degree of uncertainty in their responses before they will begin to reconsider the question and answer, this overconfidence could easily override feelings of uncertainty, and incorrect answers are left unchallenged (Gaskins, Dunn, Forte, Wood, & Riley, 1996).

Overconfidence in calibration judgments also may impact student satisfaction with academic courses and choice of academic majors. In his study of undergraduates enrolled in a macroeconomics course, Grimes (2002) found that overconfidence was linked to unmet student expectations and dissatisfaction. He concluded that for some students, "unmet performance expectations lead to dissatisfaction with the course, the instructor, and perhaps, the economics discipline in general" (p. 8). Although Grimes did not collect student satisfaction data, the argument makes intuitive sense and rings true for instructors who teach difficult or technical

subjects. Whether overconfidence and violations of expectations affect course evaluations and other indices of student satisfaction is a question that awaits further empirical study.

Underconfidence also may adversely affect student monitoring and control of comprehension and studying. Not recognizing what one does or does not understand is a failure of metacomprehension (Maki et al., 2005). That is, students may not monitor and allocate their reading or study efforts in the most efficient ways. Because students tend to be less confident on easy materials or items (Juslin, Winman, & Olsson, 2000; Lin & Zabrucky, 1998; Maki, et al., 2005), they may inappropriately devote more time than necessary to their study of material they have already mastered. In testing situations, attention and effort may be inefficiently distributed across questions and responses.

Patterns of Findings in Classroom Contexts

In this section, we describe patterns of findings from studies conducted in naturalistic classroom settings. As mentioned earlier, findings obtained in laboratory settings often provide critical insights into psychological phenomena; however, generalizing these findings to different contexts, especially classroom contexts, can sometimes be risky. Establishing strong ecological validity by generalizing laboratory findings to naturalistic classroom contexts is a different area of research and one that most often falls to educational psychologists. We have attempted to collect as much of this research on calibration in classrooms as possible, but the list may not be exhaustive and should be reconsidered representative of this line of research. Table 1 provides an overview of these studies in terms of their characteristics and major findings. These studies are discussed in the subsequent section.

Insert Table 1 about here

Achievement level and bias. As previously described, calibration accuracy has been linked to student achievement: At the global level of calibration, lower-achieving students tend to show low accuracy and overconfidence on exams, and higher-achieving students tend to show high accuracy but underconfidence (See Figure 2). This pattern has been observed among students enrolled in education, psychology, nursing, health sciences, and economics at both the graduate and undergraduate levels. The overconfidence among lower-achieving students, in particular, has been termed the "unskilled but unaware" effect (Kruger & Dunning, 1999).

Our own studies exemplify both of these effects. In both Hacker et al. (2000) and Bol and Hacker (2001), we observed this same pattern for undergraduates enrolled in an introductory educational psychology course and graduate students enrolled in a research methods course. A somewhat unexpected result in the latter study was a significant interaction between the independent variables of achievement level and item type. The calibration accuracy of higher-achieving students was similar on both multiple-choice and essay items, but the lower-achieving students were significantly less accurate on their predictions of multiple-choice items across both the midterm and final exams. These interactions did not emerge for postdiction accuracy. In Bol et al. (2005), we further replicated the findings with respect to the impact of achievement on calibration accuracy and direction of bias. This study was conducted with students enrolled in undergraduate educational foundations courses. Again, higher-achieving students were more accurate but somewhat underconfident in their predictions and postdictions than were lower-achieving students, who were largely overconfident.

Other researchers have confirmed the link between calibration accuracy and achievement. Grimes (2001) found that lower-scoring economics students were less accurate and more overconfident than their better performing peers. Similarly, Shaughnessy (1979) found a strong

positive relationship between calibration accuracy and performance on a series of four classroom exams among psychology undergraduate students. He posited that poorly performing students when judging overall performance on each exam demonstrated "an inability to distinguish adequately between known and unknown information" (p. 510). These findings were mirrored by Sinkavich (1995) who reported stronger correlations between confidence ratings and exam performance among higher-achieving compared to lower-achieving students enrolled in an undergraduate educational psychology course. Garavalia and Gredler (2002) discovered an inverse relationship between students' expected grades in an undergraduate health science course with their actual grades and GPA. Furthermore, students were divided into groups of accurate or inaccurate calibrators based on their accuracy in predicting their final grades. More accurately calibrated students who received a goal setting intervention received higher actual grades than did students who were less accurately calibrated in the control condition. This latter result supports the link between calibration accuracy and achievement.

The link between achievement level and accuracy is well established, but the relationship seems to be complicated by item difficulty. Nietfeld et al. (2005) studied undergraduate students enrolled in an educational psychology course and employed calibration measures at both global (i.e., confidence ratings on overall performance after the taking test) and local levels (i.e., confidence ratings for each item). Student performance, as measured by GPA and test scores, was a strong predictor of local accuracy and monitoring. Not unexpectedly, higher-performing students were more accurate than lower-performing students. In terms of item difficulty, students showed more accurate calibration on easy compared to difficult items. Similar to other studies that have shown the "hard-easy" effect (Flannelly, 2001; Juslin, Winman, & Olsson,

2000; Nietfeld, et al., 2005; Winne, 2004; Winne & Jamison-Noel, 2002), students displayed underconfidence on easy items but overconfidence on difficult items.

Of particular note is that only local measures of calibration were linked to student achievement levels. Nietfeld and colleagues' (2005) global measure of calibration, which was similar to the postdiction measure used in our own studies (Bol & Hacker, 2001; Bol et al., 2005; Hacker et al., 2000), did not show a relationship with achievement. Therefore, their results seemingly contradict what we had found. Our results showed a significant interaction between achievement level and item type, such that the calibration accuracy of higher-achieving students did not differ between multiple-choice and essay items, but lower-achieving students were significantly less accurate on their predictions of multiple-choice items. However, this contradiction might be explained by item difficulty. The multiple-choice items we had used (Bol & Hacker, 2001) were more difficult than essay items. Therefore, lower-achieving students, presumably with less knowledge of the tested content, should display less accuracy with difficult items. Flannelly (2001) also discovered calibration bias that varied as a function of item difficulty. In her study using undergraduate nursing students, she relied only on local confidence ratings for each test item on content related to psychiatric mental health nursing. Students' bias scores were similar on easy items regardless of achievement level but differed on difficult items.

We identified only one classroom study that did not rely on college students in their calibration research. During individual interviews, Barnett and Hixon (1997) assessed whether 2nd, 4th, and 6th graders could predict and postdict their classroom performance in spelling, math, and social studies. Overall, the students' global-level prediction accuracy was significantly correlated with achievement: High scores on the classroom tests were correlated with greater prediction accuracy. Evidence did not support uniform patterns of findings across grade levels

and subject areas. This was most likely due to the fact that the difficulty of classroom tests varied across grade levels and subject areas. For the youngest students, who faced less difficult tests than the oldest students, accuracy was quite good; however, for the oldest students tests were more difficult and their accuracy suffered. Similar to the argument that we have proposed in this chapter, Barnett and Hixon suggested that when the self-assessment capabilities of students is being investigated, the context in which it occurs must be considered.

Improving calibration accuracy. Whether calibration accuracy can be improved is a question that has not been definitively answered. Some studies have shown that improvements are difficult to obtain or are not durable (e.g., Bol & Hacker, 2001; Bol et al., 2005; Koriat, 1997; Nietfeld et al., 2005; Nietfeld & Schraw, 2002), whereas other studies have shown that various types of intervention can lead to improvements (e.g., Glenberg, Sanocki, Epstein, & Morris, 1987; Hacker et al., 2000; Nietfeld et al., 2006; Schraw et al., 1993; Yates, 1990). Three studies that were conducted in classroom contexts (Bol & Hacker, 2001; Bol et al., 2005; Nietfeld et al., 2005) demonstrated that student calibration tends to be stable despite feedback and practice.

Bol and Hacker (2001) investigated the effectiveness of using practice tests versus traditional review to improve calibration accuracy on midterm and final exams. The findings indicated that students who reviewed the content via practice tests were less accurate than students who experienced traditional review. Furthermore, calibration accuracy did not improve across exams. One explanation for the lack of improvement may be that the study included only two trials or measures, the final and midterm exam.

To address this limitation, Bol et al. (2005) investigated the impact of calibration practice on five quizzes that preceded students' predictions and postdictions on the final exam in an undergraduate educational foundations course. Feedback on quiz scores was provided

immediately to students after taking each of the online quizzes. Similar to our earlier findings, calibration accuracy on the final exam was similar for students assigned to the practice condition when compared to students who were not asked to predict and postdict their performance on the quizzes. Therefore, the practice intervention did not seem to be effective in improving calibration accuracy. Nietfeld et al. (2005) also reported that students' calibration accuracy did not improve across four course exams even though students had an opportunity to review their exam results as well as their item level confidence ratings. The authors posited that self-directed feedback, without explicit training in monitoring, was insufficient to improve accuracy.

In contrast to studies that suggest resistance to improving calibration accuracy, other experimental interventions have been successful. In some instances, the difference in results between classroom-based studies showing no change in calibration accuracy and those showing at least modest improvement may be attributable to the power or strength of the intervention.

Recently, Nietfeld and his colleagues (2006) investigated the impact of an explicit monitoring intervention on calibration accuracy, self-efficacy, and performance. Recall that in their previous study (Nietfeld et al., 2005), they failed to establish the effectiveness of repeated feedback for improving calibration accuracy and suggested that explicit training in monitoring may be necessary. Therefore, in Nietfeld et al. (2006), two sections of an undergraduate educational psychology course were randomly assigned to the monitoring and comparison groups. The monitoring intervention consisted of exercises that asked students to assess their learning for the current class session as well as their study preparation, respond to and provide confidence ratings on review items, and reflect on the accuracy of their confidence ratings. In addition to weekly feedback, the students were given feedback and interpretation on their calibration accuracy the week following the three course exams. Calibration accuracy and

performance both improved, supporting the authors' prediction that a more powerful explicit monitoring intervention is necessary to realize positive changes in accuracy.

Hacker, Bol, & Bahbahani (2007) not only studied the impact of reflection and feedback on calibration accuracy, but also the provision of extra credit points if students' predicted and postdicted scores minimally deviated from their actual scores. In their factorial design, four sections of an undergraduate educational psychology course were randomly assigned to one of four conditions: incentives and feedback, reflection and feedback, a combined treatment condition (reflection, incentives, and feedback), or a comparison condition. The reflection treatment consisted of providing students with feedback on their calibration accuracy and a questionnaire asking them to reflect on explanations for their performance, on any discrepancies between their performance and calibration judgments, and on strategies they might use to improve their calibration accuracy. We found that our intervention was successful in increasing postdiction accuracy on the last two exams for lower-achieving students in the two groups that received incentives; however, lower-achieving students in the reflection only condition were less accurate in their postdictions. There were no significant differences on measures of predictive accuracy.

Even though our reflection and feedback condition was similar to that reported by Nietfeld et al. (2006), we found contradictory results. However, different calibration measures were used in the two studies. Nietfeld relied on confidence judgments at both global and local levels (item-by-item), whereas our measures of calibration were global-level predictions and postdictions of actual test scores, not confidence judgments. Although performance, predictions/postdictions, and confidence judgments can be conceptualized as self-efficacy judgments, predictions/postdictions of performance entail other aspects of memory in addition to

self-efficacy, such as, appraisal of the memory task to be completed and translating one's ability to perform the task into a specific estimate of performance (Hertzog et al., 1990). These differences between confidence judgments and performance judgments could account for differences between the two studies, and it is up to future research to discern these differences.

Finally, findings in the Hacker et al. (2000) study illustrate the effectiveness of a complex treatment consisting of feedback, practice tests, and course instruction that included the benefits of accurate self-assessment for goal setting, time management, and academic performance. The results revealed that prediction and postdiction accuracy improved, but only for higher-achieving students. Flannelly (2001) also compared the calibration of accuracy of students who prepared for the exam by taking practice tests combined with review of the content with those that prepared via review only. Practice tests were effective in decreasing overconfidence on difficult items and underconfidence on easy items. The common element shared by these two studies that may have contributed to improved calibration was making students familiar with the type of test and test content. Thus, creating this familiarity may be a necessary condition to calibration accuracy (for a counter-example, see Bol & Hacker, 2001).

Overall, findings on the effectiveness of various interventions applied in classroom settings have yielded mixed results. It appears that feedback and practice alone are insufficient for improving calibration accuracy. With one exception (Flannery, 2001), practice tests alone do not seem to improve calibration accuracy. Reflection and instruction on self-assessment and monitoring were clearly effective in improving calibration judgment in the Nietfeld study (2006) but were found to be effective only for higher-achieving students in the Hacker et al. (2000) study. Finally, external rewards or incentives were effective in increasing the accuracy of calibration judgments only among lower-achieving students (Hacker et al., 2007).

Explanatory style. The stability of calibration accuracy demonstrated in many of the classroom studies reviewed is vexing. One would expect that students who are repeatedly provided with evidence about the inaccuracy of their calibration would modify their judgments. This does not seem to be the case. Several studies have shown that the stability of students' predictions and postdictions across multiple exams is often significantly higher than the stability of their performance (e.g., Hacker et al., 2000, Hacker et al., 2007; Schraw et al., 1993). Thus, rather than basing their calibration judgments on actual performance, past or present, which would likely be two of the best predictors of future performance, people appear to base their calibration judgments on stable persistent beliefs about their performance (Nisbett & Ross, 1980; Schraw et al., 1993). Stable beliefs about performance are encompassed under theories of explanatory or attributional style. The tendency for people to attribute failures to external causes and successes to internal causes is known as hedonic bias (Weiner, 1986) or protection of self-worth (Covington, 2004). For students, this means that they are more likely to attribute failure on an exam to external factors such as the trickiness of the items or inadequate instructor direction. Conversely, students' success on an exam is more likely to be attributed to internal causes such as the student's own ability and effort. Researchers have established links between explanatory style and metacognitive knowledge (Kurtz, Schneider, Carr, Borkowski, & Turner, 1988), which may at least partially account for the persistent stability of calibration judgments.

To investigate the potential influence of attributions on calibration, we analyzed the results obtained from an explanatory style questionnaire in our most recent studies (Bol et al., 2005; Hacker et al., 2007). Using regression analyses, we examined the unique contribution of patterns in explanatory style to prediction and postdiction accuracy on a final exam. For the outcome of prediction accuracy, we found that the more students attributed their poor calibration

accuracy to task-centered sources (external causes), the more overconfident they were in their predictions of performance. Moreover, the more students attributed their poor calibration accuracy to their own testing abilities (internal causes), the more underconfident they were in their predictions. For postdictions, only responses to the items related to task-centered (external) sources emerged as significant. The pattern observed, however, was opposite from prediction accuracy: The more students attributed their poor calibration accuracy to task-centered causes, the more underconfident they were in their postdictions (Bol et al., 2005).

The findings related to predictive accuracy seem intuitively clear because one would expect overconfidence to be associated with external explanations and lower achievement levels (e.g., "I expected to do well on the test, but the teacher wrote a terrible exam."). The findings for postdiction accuracy are more difficult to interpret. We do know that students' postdictions tend to more realistic or accurate because they have completed the exam (i.e., the "testing" effect), and they are better able to judge how they performed. One interpretation is that after completing the exam, students have a better notion of just how many items on the exam were unknown or guessed at, which, if substantial, could lead to underconfident judgments based on a perceived difficult exam.

Social Influences. Social variables influence metacognition as well as explanatory style. The influence of explanatory style in classroom contexts may be more potent due to social pressures. For example, some lower-achieving students may demonstrate a self-serving attributional style and overestimate their performance to protect their perceptions of self-worth and image of themselves as good students in comparison with their classmates.

There have been a number of studies investigating the influences of social variables on metacognition generally. For example, in a series of four laboratory studies, Karabenick (1996)

reported that the presence of co-learners' questions elicited responses reflecting cognitive dissatisfaction and feelings of confusion. Fewer studies have focused on how social variables influence calibration. Caravalho, Moisses, and Yuzawa (2001) manipulated social cues in a laboratory setting by presenting participants with information about comparative student performance from a fictitious study. Social cues had more impact on students with low versus high metacognitive ability. In a second study, they found that social cues influenced confidence judgments for only low self-regulators. The results from both studies led the authors to conclude that students with low metacognitive skills may be particularly susceptible to social influences.

We identified only two studies that investigated social influences on calibration in a classroom context. Puncochar and Fox (2004) examined undergraduate students' accuracy and confidence while cooperatively completing quizzes in small groups during class. They showed groups to be more accurate than individuals who worked alone, and that groups were more confident in their right answers. However, group confidence for wrong answers continued to increase across quizzes. The authors coined this finding as the "two heads are worse than one" effect. The effect did not diminish as a result of feedback, directions, class readings or lectures on metamemory and confidence. "Group work appears to produce the undesirable byproduct of being highly confident when wrong" (p. 590).

The second study to investigate social influences on student calibration in the classroom was conducted by Sinkavich (1995). Although the stated purpose was not to investigate social influences, the study is discussed here because the procedure clearly involved social comparisons among students on calibration accuracy. After two of the three course exams, students from two course sections were given individualized, detailed feedback on their performance and confidence ratings. In addition, they were provided with summary statistics for the class and

instructed to compare their examination feedback to their neighbors to evaluate their relative accuracy. Correlations between confidence ratings and total scores increased only for one of the two course sections and only from the second to the third exam. In the other course section, a marked decrease in prediction accuracy from the second to the third exam was observed. The author speculated that the third exam, which was a final comprehensive exam and longer than the other two, was more difficult than the earlier exams. Other findings confirmed the now familiar pattern of higher-achieving students exhibiting significantly greater calibration accuracy than their lower-achieving classmates. Sinkavich concluded that higher-achieving students were better predictors of what they do or do not know on a test, indicating better calibration accuracy. However, there was mixed support for the effectiveness of social comparisons for improving calibration judgments.

Conclusions

We introduced this chapter by adopting Nelson and Narens' (1994) guidelines for identifying our "target for research." We focused on calibration as the to-be-explained behavior, students—elementary to graduate—as the specific category of organism, and the classroom as the environmental situation. The laboratory work that has been conducted on calibration has provided many important insights into this metacognitive monitoring process, and we acknowledge that there is a continued need for such research. However, we also acknowledge that there is a need to go outside of the laboratory into more ecologically valid environmental situations. We have focused our attention on classroom applications of calibration.

There are some findings that appear to transcend context. For example, the "testing" effect appears to be salient in laboratory as well as classroom contexts: Calibration judgments made after testing tend to be more accurate than calibration judgments made prior to testing.

This seems intuitively clear in that the participants or students have much more information about the type of test, the testing items, and their performance after the test and should be able to make more accurate judgments. Also, the "hard-easy" effect is apparent in both contexts: In general, participants or students demonstrate overconfidence on difficult items but underconfidence on easy items.

However, in classroom settings, the "hard-easy" effect is compromised by achievement level. Higher-achieving students tend to be underconfident on difficult items, whereas lower-achieving students tend to be overconfident (i.e., the "unskilled but unaware" effect). Similar patterns of findings have been found in laboratory studies investigating age-related differences in calibration: Older adults as compared to younger adults tend to be overconfident in their judgments concerning subsequent recall of low-association items (e.g., Connor, Dunlosky, & Hertzog, 1997). There are obvious differences between these classroom and laboratory studies, which may make generalizations among them difficult, but there may be similar issues at stake. Perhaps variations in confidence are due to methods of calibration measurement, anchoring, or scaling effects, or perhaps underconfidence of higher-achieving students and overconfidence of lower-achieving students are the result of personal strategies used to maintain engagement in the task or to save face, respectively. Nelson and Narens (1994) argued that in laboratories, researchers attempt to control variations in participants' self-directed cognitive processing (i.e., short-circuiting via experimental control). In the classroom, however, the self-directed cognitive processing of students may provide us with much better understanding of how metacognitive monitoring is adaptively used.

Classroom investigations of calibration have shown that improving calibration accuracy is not easily accomplished. Simply providing students with practice tests and feedback on

calibration accuracy is not enough to significantly improve their accuracy. Nietfeld et al. (2005) posited that explicit training in monitoring with self-directed feedback may be necessary for improved accuracy. And, in Nietfeld et al. (2006) this was shown to be the case. This finding resonates well with the reading strategy research, which has shown the necessity for explicit training not only for monitoring strategies but control strategies to increase reading comprehension (Hacker, 2004). Other classroom results showed that improvements in calibration accuracy could be accomplished through the use of external rewards or incentives, but these appeared to be effective for only lower-achieving students. In addition, working in small groups may increase calibration accuracy, yet produce the undesirable byproduct of increasing overconfidence in wrong answers (i.e., the "two heads are worse than one" effect).

Identifying factors that contribute to calibration judgments remains a fertile area for investigation. When making local-level judgments (i.e., at the item level), students may be directly accessing their memories in search for information pertinent to the questions being asked. If memories are retrievable, high levels of confidence will be given, and more often than not, high but not perfect accuracy will result—after all, memory is fallible (for a critique of this interpretation, see Koriat, 1997).

When making global-level judgments (i.e., at the test level), the contributing factors likely become much more complex. Before a test is given, students may directly access their memories and develop an inventory of the knowledge they possess and make a prediction about their performance on a test of that knowledge. However, several of the studies we reviewed would suggest a more complicated picture. Explanatory style (i.e., the causes to which people attribute their successes and failures) accounts for a significant amount of the variance in calibration judgments, with different patterns of explanatory style being observed for higher-

versus lower-achieving students. As noted above, calibration judgments tend to be relatively stable across tasks and time. Such stability could be explained, in part, by stable personality traits, such as explanatory style. Moreover, social factors have been found to influence calibration accuracy (Caravalho, Moisses, & Yuzawa, 2001; Karabenick, 1996; Puncochar & Fox, 2004; Sinkavich, 1995). In classroom contexts, in which social influences are highly salient, finding connections between calibration accuracy and social forces would not be unexpected.

Directions for Future Research

An obvious direction for future research is to heed Nelson & Narens' (1994) advice to venture from the laboratory into the more naturalistic setting of the classroom. Given that many researchers employ convenience samples, it is not surprising that researchers tend to use their own classes. With one exception, the studies reviewed here were conducted with college students, usually enrolled in educational psychology courses. More research on student calibration across grades levels, courses, and tasks are clearly warranted. Longitudinal or cross sectional designs will help us better understand developmental changes in calibration within classroom contexts. We further endorse Nelson and Narens' position that laboratory studies are certainly beneficial when concerns about internal validity are paramount, but we also need to investigate the generlizability of these findings to the messy world of real life classrooms using authentic tasks.

As mentioned previously, student behavior in classrooms is influenced by social variables. Metacognition and calibration more specifically are no exceptions. Given the scant research examining the impact of social variables on calibration accuracy, replication studies across tasks, group compositions, and types of feedback are needed. For instance, social

comparison data in the form of calibration and performance could vary as well as the achievement level of students within groups. Lower-achieving students may benefit from social comparisons with students who demonstrate more accuracy in their calibration judgments. Such findings would be relevant to both students and teachers.

Explanatory style and other motivational variables are linked to social influences and may illuminate why calibration judgments seem to be resistant to improvement in the absence of more powerful interventions. Studies have demonstrated that feedback and practice alone are insufficient in improving calibration accuracy. This may be particularly problematic in the case of lower-achieving students who are largely overconfident. In classroom situations, lower-achieving students may be more motivated to preserve their sense of self-worth and use ego protecting strategies, such as persevering in overconfident, unrealistic predictions and relying on external attributions to explain their performance. Attributional retraining to promote more realistic metacognitive judgments, which in turn should improve monitoring ability during test preparation, represents one avenue for future study.

A final direction for future research is to augment quantitative data collection strategies with qualitative strategies in mixed methods designs. Nearly all of the classroom studies we reviewed employed quantitative designs. In our most recent study, we asked students to respond to open-ended questions to explain any discrepancies between their predictions and their actual scores. We have attempted to align these responses with findings obtained from our close-ended questionnaire assessing explanatory style related to calibration accuracy (Hacker et al., 2007). In their study on student calibration within elementary school classrooms, Barnett and Hixon (1997) relied on their analysis of classroom tests, student interviews, and classroom observations to detect patterns that may have been influenced by pedagogy, test preparation, and student

expectations across teachers, subject areas, and grade levels. Qualitative data, rich with contextual information, may direct us toward more successful interventions to improve calibration accuracy in classroom settings and ultimately improve academic achievement.

References

Allwood, C. M., & Granhag, P. A. (1999). Feelings of confidence and the realism of confidence judgments in everyday life. In P. Juslin, & H. Montgomery (Eds.), *Judgment and decision making: Neo-Brunswikian and process-tracing approaches* (pp. 123-146). Mahwah, NJ: Lawrence Erlbaum Associates.

Barnett, J. E., & Hixon, J. E. (1997). Effects of grade level and subject on student test score predictions. *The Journal of Educational Research, 90,* 170-4.

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education, 69,* 133-151.

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73,* 269-290.

Bol, L. & Nunnery, J. A. (2004). The impact of high-stakes testing on restructuring efforts in schools serving at risk students. In G. Taylor (Ed.), *In pursuit of equity and excellence: The educational testing and assessment of diverse learners* (pp. 101-117). Lewiston, New York: Edwin Mellon Press.

Broadbent, D. E. (1977). Levels, hierarchies, and the locus of control. *Quarterly Journal of Experimental Psychology, 29,* 181-201.

Caravalho, F., Moisses, K., & Yuzawa, M. (2001). The effects of social cues on the confidence judgments mediated by knowledge and self-regulation of cognition. *The Journal of Experimental Education, 69,* 325-343.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12,* 50-71.

Covington, M. V. (2004). Self-worth theory goes to college: Or do our motivation theories motivate? In D. M. McInerney & S. Van Etten (Eds.), *Big theories revisited: Research on sociocultural influences on motivation and learning* (Vol. 4, pp. 91-114). Greenwich, CT: Information Age Publishing.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5,* 69-106.

Flannelly, L. T. (2001). Using feedback to reduce students' judgment bias on test questions. *Journal of Nursing Education, 40,* 10-16.

Garavalia, L. S., & Gredler, M. E. (2002). An exploratory study of academic goal setting, achievement calibration and self-regulated learning. *Journal of Instructional Psychology, 29,* 221-230.

Gaskins, S., Dunn, L., Forte, F., & Riley, P. (1996). Student perceptions of changing answers on multiple choice questions. *Journal of Nursing Education, 35,* 88-90.

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

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

Grimes, P. W. (2002). The overconfident principles of economics students: An examination of a metacognitive skill. *The Journal of Economic Education, 33,* 15-30.

Hacker, D. J. (1998). Metacognition: Definitions and empirical foundations. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice,* (pp. 1-23). Mahwah, NJ: Erlbaum.

Hacker, D. J. (2004). Self-regulated comprehension during normal reading. In R. B. Ruddell & N. Unrau (Eds.), *Theoretical models and processes of reading, fifth edition* (pp. 775-779). Newark, DE: International Reading Association.

Hacker, D. J., Bol, L., & Bahbahani, K. (2007). Explaining Calibration Accuracy in Classroom Contexts: The Effects of Incentives, Reflection, and Explanatory Style. Manuscript under review.

Hacker, D. J., Bol, L., Horgan, D. & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92,* 160-170.

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

Juslin, P.; Winman, A., & Olsson, H. (2000). Naïve empiricism and dogmatisim in confidence research: A critical examination of the hard-easy effect. *Psychological Review, 107,* 384-396.

Karabenick, S.A. (1996). Social influences on metacognition: Effects of colearner questioning on comprehension monitoring. *Journal of Educational Psychology, 88,* 689-703.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217-273.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 297-316.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121-1134.

Kurtz, B. E., Schneider, W., Carr, M., Borkowski, J. G., & Turner, L. A. (1988). Sources of memory and metamemory development: Societal, parental, and educational influences. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory* (Vol. 2, pp. 537-542). New York: Wiley.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lin, L., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23*, 3445-391.

Lundeberg, M. A., & Fox, P. W. (1991). Do laboratory findings on text expectancy generalize to classroom outcomes? *Review of Educational Research, 61*, 94-106.

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

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

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: Volume 7, educational psychology*. New York: John Wiley & Sons, Inc.

Meyer, G. (1934). An experimental study of old and new types of examination: The effects of examination set on memory. *Journal of Educational Psychology, 25,* 641-661.

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

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

Nelson, T. O. (1999). Cognition versus metacognition. *American Psychologist, 51*, 102-116.

Nelson, T. O., & Narens, L. (1990). A theoretical framework and new findings. T*he Psychology of Learning and Motivation, 26,* 125-141.

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

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education, 74,* 7-28.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning, 1*, 159-179

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. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. The Journal of Educational Research, 95, 131-142.

Parducci, A., & Sarris, V. (1984). *Perspectives in psychological experimentations: Toward the year 2000*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pressley, M., & Ghatala, E. S. (1989). Metacognitive benefits of taking a test for children and young adolescents, *Journal of Experimental Child Psychology, 47*, 430-450.

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: Monitoring learning from text. Educational Psychologist, 25, 19-33.

Pressley, M., Levin, J. R., Ghatala, E. S., & Ahmad, M. (1987). Test monitoring in young grade school children. *Journal of Experimental Child Psychology, 43*, 96-111.

Pressley, M., Snyder, B. L., Levin, J. R., Murray, H. G., & Ghatala, E. S. (1987). Perceived readiness for examination performance (PREP) produced by initial reading of text and text containing adjunct questions. *Reading Research Quarterly, 22*, 219-236.

Puncochar, J. M., & Fox, P. W. (2004). Confidence in individual and group decision making : When “two heads are worse than one.” *Journal of Educational Psychology, 96*, 582-591.

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

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

Shaughnessy, J. J. (1979). Confidence-judgment accuracy as a predictor of test performance. *Journal of Research in Personality, 13,* 505-514.

Sinkavich, F. J. (1995). Performance and metamemory: Do students know what they don't know? *Journal of Instructional Psychology, 22,* 77-87

Weiner, B. (1986). Interpersonal and intrapersonal theories of motivation from an attributional perspective. *Educational Psychology Review, 12,* 1-14.

Weingardt, K. R., Leonesio, F J., & Loftus, E. F. (1994). Viewing eyewitness research from a metacognitive perspective. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing and knowing* (pp. 157-184). Cambridge, MA: MIT Press.

Winne, P. H. (2004). Students' calibration of knowledge and learning processes: Implications for designing powerful software learning environments. *International Journal of Educational Research, 41,* 466-488.

Winne, P. H., & Jamieson-Noel, D. L. (2002). Exploring students' calibration of self-reports about study tactics and achievement. *Contemporary Educational Psychology, 27,* 551-572.

Wright, D. B. (1996). Measuring feeling of knowing: Comment on Schraw (1995). *Applied Cognitive Psychology, 10,* 261-268.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice Hall.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zinder (Eds.), *Handbook of self-regulation* (pp. 13-39). San Diego, CA: Academic Press.

Table 1

Characteristics and Major Findings of Calibration Studies Conducted in Classroom Contexts

Study	Subjects and context	Research design	Treatment/ factors	Measures	Major findings
Barnett & Hixon (1997)	62 elementary school students in grades 2, 4, & 6 in spelling, math, and social studies	Descriptive, comparative	Grade level, subject area	Absolute, global prediction and postdiction accuracy on class assessments; scores on standardized test	Predictions more accurate in spelling and social studies than in math; no consistent grade level differences; strong correlations between calibration accuracy and achievement
Bol & Hacker (2001)	59 graduate students enrolled in 2 sections of an introductory research methods in education course	Quasi-experiment	Practice test versus traditional review for midterm and final exams; achievement level, item format	Absolute, global prediction and postdiction accuracy on course exams; achievement on course exams	Students receiving practice tests were less accurate on predictions and scored lower on multiple choice items; high achievers were better calibrated; predictive accuracy did not differ by item format for high achievers but low achievers were more accurate in their predictions of scores on essay versus multiple-choice items
Bol et al. (2005)	356 under-graduates enrolled in several sections of social and cultural foundations in education course	True experiment	Calibration practice on 5 on-line quizzes versus no quiz practice; achievement level	Absolute, global prediction and postdiction accuracy on quizzes and final exam; achievement on quizzes and final exam; explanatory style scores	No effect of the practice treatment on calibration or achievement; high achievers were better calibrated; low achievers less accurate, overconfident; explanatory style accounted for a large portion of the variance in the dependent measures.

Flannelly (2001)	66 senior year undergraduate nursing students enrolled in a psychiatric mental health course	True experiment	Practice test and feedback on confidence ratings versus no practice test or feedback, achievement; item difficulty (hard or easy)	Judgment bias (calculated by subtracting mean performance from mean confidence) on hard and easy exam items; scores on individual items	Students who received practice test with feedback exhibited less over-confidence on hard items and less under-confidence on easy items; lower achievers were over-confident but high achievers under-confident on hard items; low achievers more confident on wrong answers and less confident on right answers.
Garavalia & Gredler (2002)	69 senior year undergraduates enrolled in 2 sections of a health science course	Quasi-experiment	Goals instruction versus comparison (case study);calibration accuracy (hi versus low)	Self-efficacy for self regulated learning, goal analysis, prior achievement, final course grade	Accurate predictors who had goal setting intervention obtained higher grades than inaccurate predictors in comparison condition; inverse relationship between expected grades with actual grades and GPA.
Grimes (2001)	253 under-graduates enrolled in a principles of macroeconomics course	Descriptive, comparative	Gender; age; race; GPA, previous exposure to content; absence; study practices	Absolute, global predictive accuracy; relative global predictive accuracy (better or worse compared to first exam); exam scores	Large degree of overconfidence on both absolute and relative predictive measures; older students were less likely to over-predict performance; an inverse relationship between overconfidence and GPA; previous exposure to content resulted in greater over-predictions.

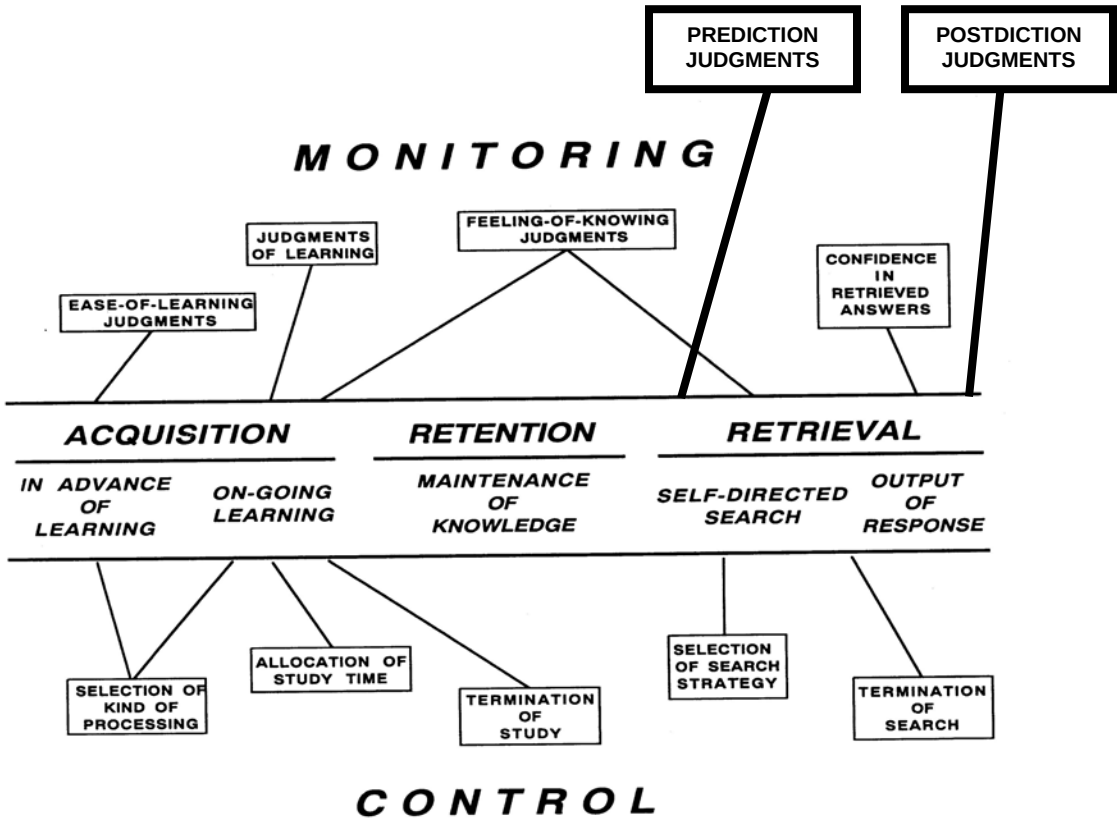
Hacker et al. (2000)	99 under-graduates enrolled in 2 sections of an introductory educational psychology course	Pre-experiment, comparative	Self-assessment instruction and practice tests, achievement level	Absolute global, predictive and postdictive accuracy, hours spent studying	Strong relationship between performance and predictive, postdictive accuracy; overconfidence among lowest scoring groups, gains in calibration accuracy among high achievers; students relied on prior calibration judgments rather then prior performance; study time was unrelated to prior performance.
Hacker el al. (2007)	137 under-graduates enrolled in 1 of 4 sections of an introductory educational psychology course	Quasi-experiment	Extrinsic incentives, reflection, both incentives and reflections, or neither; achievement level	Absolute global predictions, postdictions; predictive, postdictive accuracy; exam sores; explanatory style scores	Both extrinsic incentive conditions led to greater improved accuracy among low achievers; high achievers were more accurate calibrators; for lower achievers the explanatory style constructs predicted both predictions and postdictions.
Nietfeld et al. (2005)	27 under-graduates enrolled in an educational psychology survey course	Pre-experiment, comparative	Feedback; item difficulty, GPA	Global and local monitoring accuracy (mean difference between confidence and performance), bias scores (signed mean differences), exam scores	Monitoring remained stable over the semester; global monitoring was more accurate than local monitoring; high achieving students were more accurate in monitoring their performance; students better calibrated and underconfident on easy items but overconfident on difficult items.

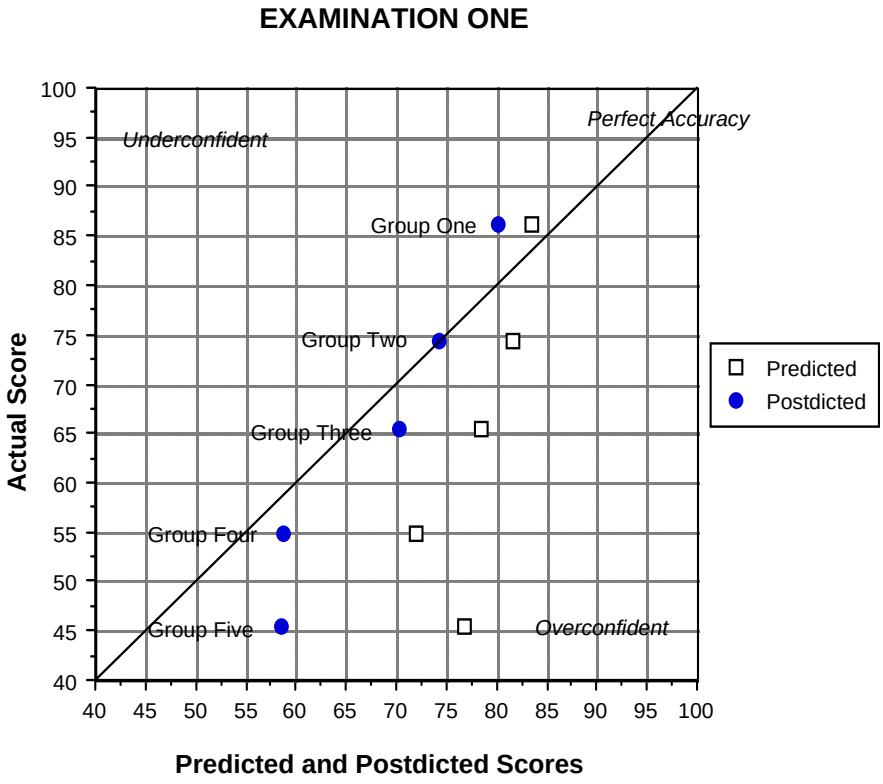
Nietfeld et al. (2006)	84 under-graduate students enrolled in 2 sections of an educational psychology survey course	Quasi-experiment	Weekly monitoring exercises and feedback vs. feedback only; gender	Local monitoring accuracy (mean difference between confidence and performance), bias scores (signed mean differences), exam and course project scores; self-efficacy	Monitoring exercises and feedback improved monitoring accuracy and performance on exams and course project; students who improved their calibration also improved their exam scores; improved calibration was associated with modest increased in self-efficacy.
Shaughnessy (1979)	47 under-graduate students enrolled in an introductory psychology course	Descriptive	Achievement levels (quartiles)	Local confidence levels (midpoint between mean on correct vs. incorrect items; confidence-judgment accuracy (ratio of local confidence over pooled variance)	Some confidence judgment accuracy even among lowest achievers; higher achievers had higher confidence-judgment accuracy scores; low achieving students were over-confident but high achieving students tended to be under-confident.
Sinkavich (1995)	67 under-graduate students enrolled in 2 sections of an educational psychology course	Pre-experiment, comparative	Extra credit for replacing incorrect with correct items on final; feedback and comparison of exam scores with classmates	Confidence ratings, exam scores	A relationship between confidence ratings and exam performance; good students had higher correlations between confidence ratings and exam performance; both good and poor students improved their scores on tests by using the replacement items

Figure Captions

Figure 1. Nelson and Narens's framework showing memory stages, examples of monitoring and control components, and the locations where prediction and postdiction judgments occur (adapted from Nelson & Narens, 1994).

Figure 2. A calibration graph plotting predicted and postdicted scores against actual scores. The calibration accuracy of each performance group can be compared against perfect calibration represented by the diagonal line (adapted from Hacker et al, 2000).


PREDICTION JUDGMENTS
POSTDICTION JUDGMENTS
MONITORING
JUDGMENTS OF LEARNING
FEELING-OF-KNOWING JUDGMENTS
CONFIDENCE IN RETRIEVED ANSWERS
EASE-OF-LEARNING JUDGMENTS
ACQUISITION
RETENTION
RETRIEVAL
IN ADVANCE OF LEARNING
ON-GOING LEARNING
MAINTENANCE OF KNOWLEDGE
SELF-DIRECTED SEARCH
OUTPUT OF RESPONSE
SELECTION OF KIND OF PROCESSING
ALLOCATION OF STUDY TIME
TERMINATION OF STUDY
SELECTION OF SEARCH STRATEGY
TERMINATION OF SEARCH
CONTROL


EXAMINATION ONE
Actual Score
Predicted and Postdicted Scores
Perfect Accuracy
Underconfident
Overconfident
Group One
Group Two
Group Three
Group Four
Group Five
Predicted
Postdicted
100
95
90
85
80
75
70
65
60
55
50
45
40
40 45 50 55 60 65 70 75 80 85 90 95 100