2000

Issues in the Measurement of Metacognition--Complete Work

James C. Impara  
*University of Nebraska-Lincoln, jimpara@unl.edu*

Linda L. Murphy  
*University of Nebraska-Lincoln*

Gregory Schraw  
*University of Nebraska-Lincoln*

Follow this and additional works at: [http://digitalcommons.unl.edu/burosmetacognition](http://digitalcommons.unl.edu/burosmetacognition)  
Part of the [Cognition and Perception Commons](http://digitalcommons.unl.edu/burosmetacognition), and the [Cognitive Psychology Commons](http://digitalcommons.unl.edu/burosmetacognition)

[http://digitalcommons.unl.edu/burosmetacognition/11](http://digitalcommons.unl.edu/burosmetacognition/11)

This Article is brought to you for free and open access by the Buros-Nebraska Series on Measurement and Testing at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Issues in the Measurement of Metacognition by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
ISSUES IN THE MEASUREMENT OF METACOGNITION
ISSUES IN THE MEASUREMENT
OF METACOGNITION

Edited by

GREGORY SCHRAW
JAMES C. IMPARA
University of Nebraska-Lincoln

BUROS INSTITUTE OF MENTAL MEASUREMENTS
University of Nebraska-Lincoln
<p>| 1.  | A Process-Oriented Model of Metacognition: Links Between Motivation and Executive Functioning | 1 |
|     | John G. Borkowski, Lorna K. S. Chan, and Nithi Muthukrishna |
| 2.  | Assessing Metacognition and Self-Regulated Learning | 43 |
|     | Paul R. Pintrich, Christopher A. Wolters, and Gail P. Baxter |
| 3.  | Assessing Metacognition in Children and Adults | 99 |
|     | Linda Baker and Lorraine C. Cerro |
| 4.  | Assessing Metacognitive Knowledge Monitoring | 147 |
|     | Sigmund Tobias and Howard Everson |
| 5.  | Metacognition and Computer-Based Testing | 223 |
|     | Gregory Schraw, Steven L. Wise, and Linda L. Roos |
| 6.  | Development of Grounded Theories of Complex Cognitive Processing: Exhaustive Within- and Between Study Analyses of Think-Aloud Data | 262 |
|     | Michael Pressley |</p>
<table>
<thead>
<tr>
<th></th>
<th>Assessing Metacognition: Implications of the Buros Symposium</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.</td>
<td>Gregory Schraw</td>
</tr>
</tbody>
</table>

Author Index 323
Subject Index 335
A Process-Oriented Model of Metacognition: Links Between Motivation and Executive Functioning

John G. Borkowski
University of Notre Dame

Lorna K. S. Chan
The Hong Kong Institute of Education

Nithi Muthukrishna
University of Natal, South Africa

The measurement of metacognition has gone through four overlapping phases: The first phase began with the insightful and stimulating paper of Kreutzer, Leonard, and Flavell (1975) on introspective reports about memory states and processes, followed by an important theoretical chapter on the nature of metamemory (Flavell & Wellman, 1977). These early contributions documented, and theoretically clarified, the fact that children could accurately report their knowledge about memory events as they related to a variety of tasks, circumstances, and strategies; furthermore, memory knowledge was shown to be age-related. A second phase quickly followed: The intention here was to show interconnections between memory knowl-
edge and memory performance. Although hindsight now reveals that a modest relationship ($r = .42$) links metamemory and memory across a wide range of learning contexts (Schneider & Bjorkland in press), an uncomfortable feeling about the “fuzziness” of the concept prevailed during this second stage of research (Wellman, 1983). From our vantage point, three interrelated conceptual and methodological problems surfaced that hindered the search for reliable and valid measures of metacognition—problems that continue to influence contemporary research and theory development:

1. Lack of clear definitions for each metacognitive construct (especially about when, where, and to whom a construct applies).
2. Lack of an array of well-analyzed tasks that permit the separation of process and performance measurements.
3. Lack of a variety of measures that converge on a given construct from multiple directions.

The third and fourth waves of research—which dominate the majority of present day studies on metacognition—focus on the issues of monitoring and control (which we refer to as executive functioning) and their associations with a variety of motivation variables. This research has been inspired, in large part, by the enthusiasm for metacognition theory, and its instructional implications for the educational reform movement. It is not surprising that current research on metacognition is more commonly found in educational psychology than in developmental psychology.

**METACOGNITION AND GOOD INFORMATION PROCESSING**

The function of metacognitive theory is to help explain successes and failures in strategy generalization. It is a theory confined principally to complex and/or novel tasks because strategies assist learners in carrying out essential cognitive operations that produce efficient, insightful learning. Strategies are at the heart of most important challenging academic activities, such as reading a difficult text passage or preparing for an examination. It is important to note that strategies are not necessarily conscious, only “potentially conscious.” Pressley, Forrest-Pressley, Elliot-Faust, and Miller (1985) have provided us with a useful definition of a strategy:

[Strategies]...are composed of cognitive operations over and above the processes that are a natural consequence of carrying out [a] task, ranging from one such operation to a sequence of interdependent operations. Strategies achieve cognitive purposes (e.g., memorizing) and are potentially conscious and controllable activities. (p. 4)
Over the last few years, the goals and prerequisites for effective strategy-based learning and instruction have been clarified by an exposition of the states and processes that comprise metacognition (Borkowski & Muthukrishna, 1992; Pressley, Borkowski, & Schneider, 1990). These goals include a clear focus on the teaching and learning of a wide variety of strategies, the higher-level processes necessary for their implementation, and the self-system (and motivational beliefs) that are their consequences as well as their sources of actualization (Borkowski, Carr, Rellinger, & Pressley, 1990). This chapter presents a process-oriented model of metacognition that is useful in understanding the ways in which strategies develop and the reasons for their generalized use over time and settings. The focus is on executive functioning and attributional beliefs, how they are conceptualized and measured, and their developmental origins.

Components of the Metacognitive System

Strategy-based learning is deliberate and effortful, at least with novice learners. It usually produces a higher level of performance than nonstrategic learning. This kind of learning is an integral aspect of what we have called Good Information Processing (Pressley et al., 1990). Although somewhere a teacher may discover a child who actually mirrors our conceptualization of the Good Information Processor, it is a rarity. Although aspects of the theory we espouse can be observed in reality, the entire model serves more as a long range goal for facilitating children’s learning through the full development of metacognitive skills than as an accurate depiction of “normal” development. Other chapters in this volume (especially those of Pintrich and Pressley) also suggest that declarative memory knowledge, memory monitoring, and cognitive self-regulation are at the heart of metacognitive theory. It is the development and integration of knowledge with higher-order skills and beliefs that are the foci of this chapter.

The unique aspect of the Good Information Processing model lies in the successful integration of the main components of the metacognitive system—including cognitive, motivational, personal, and situational characteristics. As Borkowski and Muthukrishna (1992) have argued, most of the major components of metacognition are, or can be, developed and reshaped by carefully planned classroom and home-based learning experiences—experiences that begin early and continue throughout the life-span. We have outlined 10 major characteristics that define a child who is a “Good Information Processor” (Borkowski & Muthukrishna, 1992).
1. **Knows** a large number of useful learning **strategies**.
2. **Understands** when, where and **why** these strategies are important.
3. **Selects** and **monitors** strategies wisely, and is extremely **reflective** and **planful**.
4. Adheres to an **incremental** view regarding the growth of mind.
5. **Believes** in carefully deployed **effort**.
6. Is **intrinsically motivated**, **task-oriented**, and has **mastery goals**.
7. Doesn’t **fear failure**—in fact, realizes that failure is essential for success—hence, is not **anxious** about tests—rather sees them as learning opportunities.
8. Has concrete, multiple images of “**possible-selves,”** both hoped-for and feared selves in the near and distant future.
9. **Knows** a great deal about many topics and has rapid **access** to that knowledge.
10. Has a history of being **supported** in all of the above by **parents**, **schools**, and **society at large**.

The relevant background literature and different rationales for these characteristics can be found in Ames and Archer (1987); Borkowski et al. (1990); Pressley et al. (1990); Deci and Ryan (1985); Markus and Nurius (1986); Nicholls (1984; 1989); and Pressley, Gaskins et al. (1991). Several characteristics, however, are essential aspects of our view of metacognition and deserve highlighting: (a) Strategies learned out of context, or in the rote fashion, will usually prove transient. Thus, Characteristic 2 implies that developing an in-depth awareness of how each strategy works is critical for generalized strategy usage. (b) Executive functioning is the most important process in the entire metacognitive system. Hence, Characteristic 3 emphasizes the essential role of task analysis, planfulness, and reflectivity in strategy selection as a student confronts a problem or task; the need to monitor its ongoing effectiveness; and, perhaps, to replace it with a more viable strategy. (c) Beliefs about hard work in analyzing tasks and selecting strategies as well as an orientation toward solving the task-at-hand rather than pleasing others are important motivational processes that energize self-regulatory processes. In this sense Characteristics 5 and 6 (which are motivational in nature) are related to Characteristic 3 (executive functioning or self-regulation). (d) Students need to visualize themselves in near and far time-frames in order to develop meaningful goals that will actualize the
metacognitive system at critical moments of difficulty and frustration in the course of learning and problem solving. Thus, the concept of possible selves (Characteristic 8), though understudied and not well understood, may eventually be useful in understanding why strategies are abandoned in adolescence or adulthood, in both the school and workplace. (e) Consistency in strategy instructions—across time and settings (Characteristic 10)—seems essential for lifelong strategy use to occur, for the continued development of the metacognitive system, and for the reliable and valid measurement of the components of metacognition. Some of the measurement problems encountered in this field may be due as much to inconsistencies in instruction as to the fickleness of cognitive development (cf. Siegler, 1995).

The Development of Metacognitive Theory

After outlining the major characteristics of Good Information Processing, it is useful to illustrate how these characteristics become interrelated by suggesting how the essential components of metacognition might plausibly develop. Borkowski and Muthukrishna (1992) have traced metacognitive development in terms of what happens to a child who receives high quality, interactive strategy instruction in both the home and school:

1. The child is initially taught to use a learning strategy and, with repetition, comes to learn about the attributes of that strategy (this is called specific strategy knowledge). These attributes include the effectiveness of the strategy, the range of its appropriate applications, and how to use it with a variety of tasks. Figure 1 shows how a simple strategy (such as summarization), in isolation from the rest of the system, can be expected to produce an improvement in performance.

2. Next, the child learns other strategies and repeats them in multiple contexts. In this way, specific strategy knowledge is enlarged and enriched. Figure 2 presents a schematic diagram showing the emergence of a number of specific strategies. The child comes to understand when, where, and how to deploy each strategy.

3. The child gradually develops the capacity to select strategies appropriate for some tasks (but not others), and to fill in knowledge gaps by monitoring performance, especially when essential strategy components have not been adequately taught. At this stage, higher-order executive pro-
Figure 1. A primitive view of the strategy use-performance relationship.

Figure 2. Multiple strategies and performance.
cesses emerge. This is the beginning of self-regulation, the basis for adaptive, planful learning and thinking. Figure 3 shows the relationship of executive processes to specific strategies. Initially, the function of the executive is to analyze the task at hand and to select an appropriate strategy; during the course of learning, its role shifts to strategy monitoring and revision.

Figure 3. Executive functioning and strategy use.

4. As strategic and executive processes become refined, the child comes to recognize the utility and importance of being strategic (general strategy knowledge accumulates), and beliefs about self-efficacy develop. In addition, as the child acquires domain-specific knowledge and skills, beliefs about efficacy become differentiated across domains. More specifically, children learn to attribute successful (and unsuccessful) learning outcomes to effort expended in strategy deployment rather than to luck or to task difficulty encountered in specific domains of study. Furthermore, some children come to understand that through self-directed actions mental competencies can be enhanced.
In these ways, the metacognitive model integrates cognitive acts (in the form of strategy use) with their motivational causes and consequences. Figure 4 suggests that following most cognitive acts, the child is often provided with, or infers, feedback about the correctness of performance and its specific cause(s). This feedback is essential for shaping personal-motivational states (e.g., attributional beliefs, which in turn can energize the executive processes necessary for strategy selection and deployment in future situations.

5. A sense of self-efficacy and an enjoyment of learning flow from individual strategic events and eventually return to energize strategy selection and monitoring decisions (i.e., executive processes). It is this latter connection—the association between the learner's reasons for learning and the deployment of self-regulation—that has been absent from most instructional programs. This theme is at the heart of our most recent extensions of metacognitive theory (Borkowski et al., 1990; Borkowski & Muthukrishna, 1992).

---

**Figure 4.** Motivational correlates and causes of strategy use.
6. General knowledge about the world as well as domain-specific knowledge (e.g., math) accumulate. Such knowledge is often sufficient to solve problems, even without the aid of strategies. In these situations, metacognitive processes, such as strategy selection are unnecessary, although some motivational components may remain functional and important (see Figure 5).

7. Crystallized visions into the future help the child form a number of “hoped-for and “feared” possible-selves (Markus & Nurius, 1986) providing the impetus for achieving important short-term as well as long-term goals, such as becoming a “competent student” in order to eventually become a “successful lawyer” (cf. Day, Borkowski, Dietmayer, Howsepian, & Saenz, 1992). In this way the self-system takes on a futuristic perspective, providing goals and incentives that stimulate the operation of the entire metacognitive system. The complete metacognitive model, including the self-system and the domain-specific knowledge “bypass,” is presented in Figure 5.

In summary, the centerpiece of metacognitive theory is strategy selection and use. Not only are specific strategies essential for effective learning and problem solving, they also provide the context for training higher-level planning and executive skills explicitly as well as represent the basis for restructuring attributional beliefs and enhancing self-efficacy. As such connections are formed and ingrained, instructional emphasis can shift to their interface with domain-specific knowledge and the explicit incorporation of possible-selves training into individualized curricula. It is hoped that the net result of integrating and instructing these central and peripheral components of metacognition will be the production of more effective and efficient students, who share many of the characteristics of the “Good Information Processor” (Pressley et al., 1990)

A Test of the Model

*Measurement approaches.* There are three general approaches that have been used to manipulate and/or measure components in the metacognitive model during the first two decades of research on this topic:

1. Set up conditions in which no other processes appear as reasonable, alternative theoretical explanations.
2. Instruct processes directly (and hope that “nothing else” has been trained).
Figure 5. Cognitive, motivational, and self-system components of metacognition: The complete model.

3. Develop a broad-based (or domain-specific) questionnaire that reflects students' use of (or beliefs about) the attributes of a metacognitive state or process and relate individual differences to performance.

Several points of clarification about these measurement approaches are in order. First, they need not be mutually exclusive; for instance,
it is possible (and desirable) to train and assess processes and beliefs within the same study (i.e., Parts 2 and 3 combined). Second, the first two approaches demand a theoretical respect for a clear distinction between process and performance as they relate to metacognitive measurement. In a seminal paper, Belmont and Butterfield (1977) argued that by measuring performance, separate from the processes from which it presumably flows, research on cognitive development stands on firmer theoretical ground, especially when inferring the former from the latter. Third, although we have listed only three historically rooted research approaches to measurement, the new technique advocated by Pressley (this volume)—think aloud, protocol analysis—represents a powerful context in which to observe and measure metacognitive activity as it is occurring.

At times, metacognitive research has utilized all three measurement approaches. From our vantage point, this style of research is a particularly powerful way to validate metacognitive models, especially if the combination of approaches results in internal replication. In our own research program, a study by Reid and Borkowski (1987) contains aspects of all three methodologies, especially the latter two (process manipulations and questionnaires designed to assess changes in performance, strategy use, attributional beliefs, and cognitive styles following a multi-faceted strategy-based intervention).

Before describing the Reid and Borkowski (1987) study in detail, it should be noted that there are relatively few studies where the researchers have tried to assess how the major components of metacognition interrelate. The reason is that it is difficult to manipulate, or observe, metacognitive components in isolation from one another. This is an important point for measurement in this area. It is also the case that theoretically distinct components may not be entirely separate from one another as they operate in the real world: It is often easier to develop theories with boundaries and boxes than to locate, isolate, and measure these same processes in laboratory or observational settings. In order to be assessed reliably, components of metacognition may need to be measured in the midst of their complex interactions, rather than in isolation.

An integrated approach. In an early study of interrelationships among the components of metacognition, Reid and Borkowski (1987) attempted to establish the plausibility of the metacognitive model with children who were learning disabled. The unique and combined effects of training specific strategy knowledge, teaching self-control skills, and reshaping attributional beliefs about the importance of effort were studied. More specifically, three treatment groups were
compared: a self-control condition, a self-control plus attributions condition, and a control condition. In the self-control condition, the teacher modeled self-verbalization procedures for the child (e.g., “look to see how the problem might be solved; stop and think before responding”). These self-control procedures were taught in the context of specific strategy training, which focused on the use of interrogative-associative mediators appropriate for a paired associate task and a clustering-rehearsal strategy for use on a sort-recall readiness task.

In the self-control plus attributions condition, children received strategy and self-control instruction as well as attributional training designed to enhance both antecedent and program-generated self-attributions. Antecedent attribution training took the form of a discussion focusing on general, pervasive beliefs about the causes of success and failure; children were also given opportunities to perform previously failed items in the self-control package. Program-generated attributions consisted of feedback about the relationship between strategic behavior (or its absence) and performance during paired-associate learning. Individual items were shown to be correct or incorrect depending upon whether effort was put forth in deploying the appropriate strategy. The control group received the same amount of strategy training as the experimental groups but did not receive self-control or attributional training.

Widespread strategy generalization occurred on a 3-week posttest in the self-control plus attributions condition. More importantly, the persistent use of strategies was maintained at a 10-month follow-up. In addition, attributional beliefs and metamemory were permanently altered in this condition. These results seem surprising in light of the longstanding difficulties in obtaining strategy generalization. For example, Gelzheiser (1984) was unable to obtain extensive generalization in learning-disabled children following prolonged training; attributional retraining, however, was not a component in her instructional package. We believe that the emphasis on strategy-based effort set in motion a bidirectional chain of events between strategic acts and the growth of positive beliefs about the importance of effort in deploying strategies. The net result was that children who, for the most part, were not spontaneous strategy users at the study’s outset, deployed strategies with greater flexibility and persistence up to 10 months following the end of training.

The intervention program in the Reid and Borkowski (1987) study contained three key components: detailed information about two specific strategies, self-control procedures useful in implementing
these strategies, and an explicit recognition of the role of effort and personal causality in producing successful performance. The interaction of these metacognitive components seemed to play an essential role in the generalization of strategic behaviors. These results, together with those of Borkowski, Weyhing, and Carr (1988) and Carr and Borkowski (1989) on the explicit training of attributional beliefs in the context of reading comprehension instructions, lead us to believe that long-term changes in strategic behaviors are probably dependent on the development of complex relationships among components in specific strategy knowledge, self-regulation, and motivational beliefs. In a sense, this set of studies has expanded the boundaries of cognitively based interventions by focusing on how self-regulation, the heart of metacognition, depends on children's rationales and attitudes about the learning process per se and how they conjointly contribute to academic achievement.

In subsequent sections, we trace more recent advances in the theory and measurement of executive functions and attributional beliefs. Finally, we suggest specific contexts that influence the integrated development of metacognition with a view toward understanding more about situational factors related to when and where metacognitive measures should best be gathered.

THEORIES OF EXECUTIVE FUNCTION

Although the major components of executive functioning are by no means agreed upon, most researchers would concur that the three components represented in Figure 6 are essential. The first, and perhaps most essential, component, is *task analysis*. Despite its centrality in defining executive processing, it is the most poorly understood, and least often measured process in the system. The importance of task analysis lies in its potential for explaining generality across settings and domains. This aspect of the executive is critical because its proper execution is essential for the occurrence of the second activity—*strategy selections*. A related component—*strategy revision*—is closely linked to strategy selection and is observed on tasks that allow for the measurement of continuous changes in the processes that determine successful performance in the face of changing task demands. It is probably methodologically easier—and perhaps theoretically wiser—to measure strategy revisions than initial strategies selection in that "moments" of strategy change are likely to be more reliably assessed than strategy initiation (Siegler, 1995). The most widely studied attribute of executive processing is *strategy monitoring,*
which has a long and substantial history in developmental, educational, and cognitive psychology (Borkowski, Milstead, & Hale, 1988; Schneider & Pressley, 1997). Pintrich (this volume) does an excellent job of classifying the types of monitoring tasks that have been used in metacognitive research, and Schraw, Dunkle, Bendixen, and Roedel (1995) have recently suggested that monitoring skills are often domain general. We turn now to a review of several theoretical positions that describe the interrelationships among, and the functioning of, the major components of executive functioning as well as their connections with other aspects of cognitive systems.

Figure 6. Major attributes of executive functioning.

**Components of Executive Functioning**

- Task Analysis
- Strategy Control (Selection & Revision)
- Strategy Monitoring
Butterfield's Theory of Executive Functioning

Butterfield, Albertson, and Johnston (1995) have developed a new theory of cognition in which executive functioning plays a critical role. In their model, cognition, metacognition, and executive functioning are three major components. The cognitive level consists of all the knowledge and strategies that exist in long-term memory; this reservoir of information about the cognitive system is critical for effective problem solving. The metacognitive level represents awareness of the cognitive level and contains "models" of the various cognitive processes as well as an understanding of how knowledge and strategies interconnect. This level is the unique aspect of the Butterfield et al. (1995) theory in that it rests on the interesting assumption that metacognitive skills are generalizable—but only if students develop mental models in their cognition system. Furthermore, the metacognitive level is potentially trainable.

Executive functioning coordinates the two levels—the cognitive and the metacognitive—by monitoring and controlling the use of the knowledge and strategies in concordance with the "mental model building." Thus, in the Butterfield et al. (1995) theory, in contrast with the theory of Day, Borkowski et al. (1992) described earlier, the metacognitive level is distinct from the mechanisms that help to control and monitor the cognitive level. For Butterfield these mechanisms seem to represent executive functioning in operation.

The concepts of monitoring and control, which are responsible for the emergence of complete and mature mental models, allow for the possibility of a more general theory of cognition than has previous task-specific theories. Butterfield et al. (1995) believe that individuals are able to create mental models about their own cognitions based on their day-to-day problem solving activities. They suggest that these models are similar to those developed by scientists through prolonged, detailed task analysis. Self-generated models exist in direct relation to the knowledge and strategies present at the cognitive level. The development and integration of task-specific models, made possible by executive functioning, eventually lead to a personalized (and unified) theory of cognition. Individuals who possess such unified theories, according to Butterfield and Albertson (1995), should show more rapid acquisition and more extensive generalization of skills and strategies across domains.

Bransford's Ideal Problem Solver

Bransford and Stein (1993) have incorporated aspects of executive functioning into their model of the IDEAL problem-solver. The
acronym, IDEAL, is used to symbolize the skilled components in problem-solving: (a) **Identify** an important problem to-be-solved; (b) **Define** the subgoals involved in solving the problem; (c) **Explore** possible approaches to the problem, that is, select a set of potential strategies; (d) **Anticipate** potential outcomes before acting on the best initial approach; and (e) **Look** back and learn from the entire problem-solving experience. Because these five steps are used flexibly by expert problem-solvers, they do not always occur in the same fixed order nor is each step necessary for all problem-solving tasks.

These five steps, proposed by Bransford and his colleagues, closely resemble the components of executive functioning discussed earlier. The first steps—problem identification and definition—represent a form of task analysis. The discovery and definition of an existing problem shape the next steps—exploring approaches and anticipating outcomes. In these steps, various strategies are considered and the best alternative is chosen. The last step of the IDEAL problem solving strategy involves looking back and learning from prior efforts. In the ongoing process of problem solution, this step is at the heart of what we have called strategy monitoring and revision.

Bransford and his colleagues have incorporated aspects of the IDEAL problem-solver into their video-based technology research. The Cognition and Technology Group at Vanderbilt has developed a technology that anchors and situates instruction in shared environments, thus permitting sustained exploration by students and teachers (Bransford, Sherwood, Hassebring, Kinzer, & Williams, 1990). Students experience the value of exploring the same setting from multiple perspectives (e.g., as a scientist, historian, and mathematician). As they discover their own issues to explore in these enriched environments, they communicate their ideas to other students and develop analytic skills as a result of their problem-solving activities. Difficult to discern in the research of the Vanderbilt group are the precise sets of metacognitive skills that emerge as a result of these shared experiences, and their reliable measurement, as students acquire prolonged experience with video-based instruction. What specific problem-solving strategies are developed? Are higher-level planning, task analytic, or monitoring skills (i.e., executive functioning) enhanced? Are specific beliefs about self-efficacy and the personal challenge to develop one’s own mind explicitly fostered?

We suspect that a comprehensive, and carefully used, video-technology approach to instruction influences the emergence of planning and executive skills as well as enhances motivational beliefs about self-efficacy. More precise assessment of these characteristics
would help to advance the metacognitive aspects of the theories that underlie video-technology. It is to the measurement of personal beliefs, and other motivational states, that we now turn.

**ATTRIBUTIONAL BELIEFS AND METACOGNITION**

An important component of the personal-motivational states in the metacognitive model is what students perceive as the causes of their successes and failures in school. The most common reasons students give for their successes and failures are ability, effort, their attitude (such as interest), physical factors (mood, fatigue, etc.), task difficulty, assistance from others, and luck. Weiner (1983, 1984) has classified these attributions as either internal or external locus, constant or variable over time and across different situations, and controllable or uncontrollable by oneself. For example, ability attributions have an internal locus, are stable but uncontrollable whereas effort attributions have an internal locus, are unstable (therefore can be changed) but are controllable. Each of these dimensions is proposed to be uniquely associated with particular psychological consequences. The locus dimension affects self-esteem (e.g., attributing success to internal factors increases self-esteem). The stability dimension relates to changes in expectancy of success or failure and affective reactions (e.g., attributing failure to a stable cause such as lack of ability leads to high expectancy of future failure and hence feelings of hopelessness). The controllability dimension relates to sentiments and evaluations of others (e.g., if a student fails because of a controllable cause, such as lack of effort, anger is often elicited and the student is negatively evaluated). Affective reactions and anticipations in conjunction with expectancy of success are assumed to affect a student’s willingness to try, persistence, choice or avoidance of tasks, and, eventually, task performance.

Research has indicated that students who attribute their successes and failures in school tasks to internal and controllable sources (e.g., one’s own effort) are more likely to persist in the face of difficulty (Nicholls, 1984; Weiner, 1984). If students are convinced that success or failure depends on effort, they will realize that they can expect success if they put in the required effort. These students who have internal perceptions of control have high expectancy of success and are motivated to work hard because they realize that success or failure will depend on their own effort. On the other hand, students who attribute successes and failures to external or uncontrollable sources (e.g., powerful others, luck, task difficulty, or inherent abilities) are
more likely to give up when they come across difficulties in their learning. Students who attribute school success to luck will not be confident of maintaining that success at all times and will not be motivated to expend maximum effort to attain prescribed learning goals. Likewise, students who think that their progress in school depends entirely on teachers' skills will not be motivated to become independent in learning. Furthermore, they will not be motivated to try hard because they do not see that their effort will contribute to success.

It has been widely accepted that beliefs in personal control over task outcomes can be promoted by convincing students that school successes and failures are attributable to effort. Such an approach has not been entirely successful. Some students, particularly students with learning difficulties, may find that they keep on failing in spite of increased effort, particularly if they do not know how to try harder. Such negative experience would even further reinforce their beliefs in the lack of ability, and thus increase feelings of helplessness. Probably a more fruitful direction is to try getting these students to attribute failures to both insufficient effort and ineffective task analysis (Borkowski, Weyhing, & Turner, 1986; Clifford, 1986; Licht & Kistner, 1986). There are many advantages of encouraging strategy attributions in students, including the elimination of the guilt associated with not trying hard or the embarrassment and public shame associated with being stupid. More importantly, strategy attributions allow failure outcomes to be seen as problem-solving situations in which the search for a more effective strategy becomes the goal (Clifford, 1986). Indeed, effort and strategy attributions play a critical role in the developmental aspects of metacognitive theory.

Role of Attributional Beliefs in Metacognitive Theory

As discussed earlier, the centerpiece of metacognitive theory is strategy selection and use—that is, the operation of executive function in the form of self-regulation. It was explained in the previous section that executive functioning is responsible for the planning, selecting strategies, monitoring, evaluating, and revising ongoing performance in learning and problem solving. Such planning, evaluating, and regulating processes require effort, initiation, and willingness to try, as well as persistence. Furthermore, there needs to be some minimal expectancy of success before a student is prepared to try, marshal the appropriate effort, and persist when encountering difficulties. If there is little or no expectancy of success, students will likely expend little
effort in learning, or they may even actively avoid tasks that they perceive will eventuate in failure. Hence, the assessment of attributional states likely represents an essential step in measuring any aspect of executive functioning.

Before students are prepared to deploy effort in planning, evaluating and regulating strategy use, they must develop and maintain four beliefs:

1. The value of good performance on the task at hand: That is, they must want to do well and strive to obtain a good result;
2. Personal control over task outcomes: That is, they must be convinced that success or failure on the task depends on themselves;
3. Usefulness of strategy use: That is, they must have the knowledge that use of specific strategies will lead to better performance on the task;
4. Their ability to use strategies effectively and successfully: That is, they must perceive themselves as capable and competent.

In other words, students who are committed to do well on a given task, who have well-developed specific strategy knowledge, and who believe that their effortful use of strategies will lead to successful task performance are likely to be active in strategy selection, monitoring, and regulation. Empirical support for these theoretical propositions is starting to emerge. For instance, perceptions of personal control (effort and strategy attributions) have been shown to relate positively to knowledge and use of strategies, and to academic performance (Borkowski, Weyhing, & Carr, 1988; Chan, 1994; Chan, 1996a). Further, the pattern and impact of attributional beliefs appear to change across the school years (Clayton-Jones et al., 1992).

In the Clayton-Jones et al. study, students from grades 4, 6, 7, 9, and 11 were administered a general attribution scale incorporating ability, luck, effort, and strategy attributions for success and failure. For the primary grade children, effort attribution for success was positively related to achievement in Math and English (a combined score) but at grade 9, strategy attribution for success emerged as a positive predictor of achievement. Ability attribution for failure, however, was a pervasive negative influence across all grades.

The positive effects of beliefs in personal control over task outcomes on the use of strategies were also observed in both gifted and average ability students in grade 7 (Chan, 1996a). Indeed, the relationship between attributional beliefs and use of strategies was fur-
ther clarified in the Chan (1994) study involving 104 grade 5, 133 grade 7 and 101 grade 9 students. Path analysis results indicated that students in the higher grades (7 and 9) who believed that they had personal control over learning outcomes, who were not inclined to feel helpless in their learning, who had high self-perceptions of cognitive competence, and who had good knowledge of strategies, were more likely to use strategies in their learning. For grade 5 students, however, only the perceived competence measure was found to influence use of strategies. When reading achievement was included, hierarchical regression analyses revealed that although the attributional beliefs and perceived competence had a more important role (relative to strategic learning) in explaining reading achievement in the younger grades, in grade 9 the role of the strategy knowledge and usage variables was as important, if not more important, as the motivation variables. Path analyses results clarified these relationships: Knowledge and use of strategies were found to mediate between the effects of attributional beliefs and perceived competence on reading achievement for grade 9, but not for the younger grades. Results of the grade comparisons indicated that strategy attributions were not prominent in students' attributional beliefs before grade 9. This result could explain the lack of influence of strategy knowledge and usage on reading achievement in the younger students.

Assessment of Attributional Beliefs

The findings of the research studies described above highlight the complex relationships between the various components of metacognition and their developmental differences. It follows that to advance our knowledge and understanding of the development of metacognition, the components of metacognition should be studied as they interact with each other in specific learning contexts and from a developmental perspective. The study of students' attributional beliefs and their impact on the executive processes and academic performance provides a useful example to illustrate this principle. This entails as a starting point the search for effective means for obtaining information on attributional beliefs. This is not an easy task because students themselves are not fully conscious of the existence of learned helplessness or control beliefs, or they may encounter difficulty in reporting their causal attributions. We now turn to some of the issues in the assessment of attributional beliefs that need to be addressed.

Assessment method. Earlier research on causal attributions in school learning tended to measure attributional beliefs by requiring
respondents to choose a single major cause. The resultant attribution was then classified as internal or external, stable or unstable, controllable or uncontrollable according to Weiner’s (1984) classification scheme; inferences were then drawn as to the likely psychological and behavioral consequences. However, the unquestioned acceptance of the categorization often causes confusion as it was often the perceived stability from the perspective of the respondent, rather than the stability implied by the objective task characteristics, that was the determinant of the affective outcomes (Weiner, 1983; 1984). To measure attributional beliefs, respondents could be asked to rate the cause in question on the stability or controllability dimensions direct rather than using an a priori classification of the causes. However, the differential consequences of the various combinations of locus, stability, and controllability dimensions complicates such an approach. This is particularly so when the dimension of intentionality is subsequently added (Weiner, 1984).

Elig and Frieze (1979) compared different methods of assessing causes of success and failure, including open-ended questions (e.g., why do you think you succeeded on this task?), independent unipolar ratings (e.g., rate each given cause on a 5-point scale), ipsative measures such as percentage assessment (e.g., provide a percentage contribution for each given cause), choice of one cause (select one from a given set), bipolar ratings (rate each of two causes that are different on a particular dimension), and paired comparison (from among several causes). Results indicated that the independent unipolar rating method was the superior technique as it had good face validity, did not force intercorrelations among attributions, and had moderately good intermethod correlations with percentage measures.

**Strategy use as a distinct attribution.** As yet little research has been done in the development of attributional beliefs with respect to the use of strategies. Most of the extant work has focused on attributions to ability versus effort (e.g., Cooley & Ayres, 1988; Kistner, Osborne, & LeVerrier, 1988; Wigfield, 1988). Given the critical role played by effort and strategy attributions in energizing the executive processes in the development of metacognition, we need to extend our current knowledge on the development of strategy-related attributions.

Research findings have indicated that children’s concepts of ability become differentiated with age (Nicholls, 1978; Nicholls & Miller, 1984). From a review of research, Stipek and Maclver (1989) concluded that children in preschool and early elementary school have a global concept of ability that includes social behavior, work habits, and conduct, and that they conceptualize ability as an “instrumental-
incremental" skill that is increased by practice and effort. Over the primary school years (third to sixth grade), children’s definitions of intellectual ability become narrower and the concept of ability as a stable trait emerges. However, it is not until early adolescence that they fully differentiate ability from effort and conceptualize ability as an “entity” unaffected by effort. Nicholls (1978) suggested that this mature concept of ability as a stable trait, unaffected by effort, requires an understanding of the reciprocal relationship between effort and ability—that ability limits the effectiveness of effort and that effort is more facilitative of performance in high-ability than in low-ability individuals. Clearly, some form of formal operational thought is necessary for this understanding to emerge. Apart from cognitive development, systematic changes in the activities, organization, evaluation practices, and ability-grouping patterns that children are exposed to in school may also contribute to developmental shifts in children’s ability judgments (Stipek & MacIver, 1989).

Likewise, the differentiation of the concept of strategy use from effort may also be age-related, particularly because strategy instruction has not been given much emphasis, at least until recently. It is critical to find out when strategy attributions becomes prominent in students’ motivational orientations. Some evidence is emerging from the Chan (1994) study described earlier, suggesting that the differentiations between ability, effort, and strategy attributions may not occur fully until the high school years. Furthermore, data from a recent cross-sectional project (Chan & Moore, 1994) gave support to the distinctiveness of strategy attributions as separate from ability, effort, and luck attributions.

**Subject-specificity.** Most of the research on causal attributions has been limited to general notions of learning rather than learning in specific subject domains. Marsh, Cairns, Relich, Barnes, and Debus (1984), however, maintained that there is good evidence for the separation of attributions according to academic subject matters, at least in the case of ability attributions. The results of their study suggest that attributional responses students make do not generalize across academic subject domains and two subject-specific dimensions (ability in mathematics and reading) can be identified. It was suggested that ability attributions are specific to academic content, but effort attributions and external attributions may not be subject-specific. Similarly, strategy attributions may also depend on specific subject domains. These findings suggest that students held different attributional beliefs for different school subjects. Such research provides rather compelling evidence for moving to subject-specific as-
sessments as well as global assessments of motivation and strategic knowledge.

Development of a Causal Attribution Scale

A 10-item rating scale was developed and used in several research projects (e.g., Chan, 1994; Chan, 1996a, 1996b; Clayton-Jones et al., 1992) to assess students' tendency to attribute their school success and failure experiences to the four likely reasons of effort, ability, strategy use, and luck. Five items describe success incidents (such as doing well on a test), and the other 5 describe failure incidents. For each item, four different reasons are listed and students are required to rate each on a 4-point scale to indicate how true they consider that particular reason to be for them. Two versions were constructed, one for high school students and one for primary students. The content in the versions was the same, only the wording was modified to suit the students' grade level. The following is a sample item from the high school version:

<table>
<thead>
<tr>
<th>Rarely True</th>
<th>Sometimes True</th>
<th>Often True</th>
<th>Almost True</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. When you received a bad school report, it was probably because</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. you aren't very good at schoolwork</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>b. you didn't try very hard</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>c. you didn't have any useful methods for studying</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>d. you were having a lot of bad luck at the time</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Each of the four types of reasons (effort, ability, strategy use, and luck) is grouped across the five success items and the five failure items, respectively, thus yielding eight separate subscales. For example, a high score on the Failure-Ability subscale indicates greater tendency of attributing school failure experiences to a lack of ability. Based on this general version, three other scales were subsequently developed, one for English/Reading, one for Mathematics and one for Social Studies. In the subject-specific scales, the subject area (e.g., math) was specified or inserted in place of expressions like "schoolwork." Again, two versions—for primary and high school students—were developed for each subject area, thus giving a total of eight causal attribution scales.
As part of a 3-year longitudinal study, Chan and Moore (1994) administered these scales to 354 students in grade 5, 650 in grade 7, and 450 in grade 9. The data from the Causal Attribution Scales were subjected to several Confirmatory Factor analyses using the LISREL (Joreskog & Sorbom, 1989) to examine the invariance of factor patterns across grades and across subject areas. The use of Confirmatory Factor Analysis allows the fit of a hypothesized a priori factor pattern, indicating which items should load onto which factors, to be tested against the empirical data. The analysis provides goodness of fit statistics, which indicate how closely a matrix obtained from parameter estimates for the posited model correspond to the input correlation or covariance matrix calculated from the data. Each set of four subscales (effort, strategy, ability and luck attributions) for success and failure for the three grade levels, as well as the combined total group, was analyzed separately. Overall, results confirmed a four-factor pattern in each case, with the items loading clearly on the intended factors. At the same time, there was some indication that the distinctiveness of the strategy attribution from the effort and ability attributions increases with age. This was seen in the decrease in the factor correlations and in the cross-loading of the factor score regressions from grade 5 to 9.

A preliminary analysis of data from the general and subject-specific scales revealed only moderate correlations between the general and the English, Mathematics and Social Studies scales. The correlations ranged from .55 to .61 for ability attributions, .56 to .65 for strategy attributions, .64 to .79 for effort attributions and .60 to .62 for luck attributions. Furthermore, ANOVA results revealed subject-domain differences as a function of grade level for ability and effort attributions, independent domain and grade level differences for strategy attributions, but no differences for luck attributions (Moore & Chan, 1995). For example, students were more likely to make ability attributions for failures in specific subject-domains than in the global domain, and the younger students were more likely to make ability attributions for successes in English/Reading than in the other domains. Whereas younger students were more likely to make effort attributions for successes in English/Reading and Social Studies than in Maths and the global domain, the subject-domain differences observed among the older students were in the reverse direction: They were more likely to make effort attributions for successes in Maths and in the global domain than in English and Social Studies. For strategy attributions, students were more likely to attribute failures in specific subject domains rather than the global domain to their
lack of effective strategy use, whereas the reverse was observed for successes. No grade level nor subject-domain differences were observed for luck attributions. These findings once again illustrate the need to consider contextual and developmental differences in any study of metacognition and its components.

Pattern of Attributional Beliefs: Adaptive versus Maladaptive

In interpreting the scores from the Causal Attribution Scale, and to make inferences as to the consequences of particular beliefs, we need to examine the pattern of a student’s tendency to attribute success or failure to ability, effort, strategy, or luck. The likelihood of making any one of these attributions by itself is not sufficient to allow meaningful assessment of the components of metacognition as they interact. To minimize the number of measures to be included in an analysis as well as to facilitate interpretation of the results, the eight attribution subscale scores can be combined to form two or more variables (e.g., a “belief in personal control” variable, which can be the mean of ability, effort, and strategy attributions for success and effort and strategy attributions for failure; and a “learned helplessness” variable, which can be the mean of luck attribution for success and ability attribution for failure).

In the previously noted Chan (1994) study using students from grades 5, 7, and 9, significant differences were observed between students with and without learning difficulties (LD) on such patterns of adaptive versus maladaptive attributional beliefs. LD students were more likely than the non-LD group to attribute successes to luck and failures to lack of ability or bad luck, but less likely to attribute successes to effort or use of effective strategies. That is to say, compared to non-LD students, LD students had greater maladaptive learned helplessness beliefs, but less adaptive control beliefs. When these adaptive and maladaptive composite scores were used in path analyses instead of individual subscale/attribution scores, a consistent trend started to emerge. Maladaptive attributional beliefs tended to have a direct negative influence on performance/achievement, whereas the positive influence of adaptive attributional beliefs on performance was consistently mediated through knowledge and the use of strategies. These relationships were observed in school-age students (Chan, 1994, in press-a; Ee & Chan, 1994; Youlden & Chan, 1994) as well as in university nursing students (Cholowski & Chan, 1994). It seems likely that although maladaptive attributional beliefs may have a direct detrimental effect on performance, adaptive
attributional beliefs are not sufficient to bring about better performance on their own: Each student must also have good knowledge and effective use of strategies. Adaptive attributional beliefs, or beliefs in personal control over task outcomes, serve the function of energizing the executive processes responsible for the regulation of strategies and, in combination, are likely to lead to better classroom- or laboratory-based performance.

Suggestions About Measuring Metacognitive Components

In the initial wave of research, the components of metacognition were measured in isolation. For instance, in the Kurtz and Borkowski (1984) study, knowledge about a set of memory problems reported by impulsive and reflective children was related to their transfer of reading strategies 3 years later; no intervening changes in other aspects of metacognition, such as the development of monitoring or control skills, that might have been associated with prior metamemorial knowledge, and perhaps causally related to the development of reading strategies, were assessed. This study illustrates the need to consider (and perhaps control) multiple aspects of metacognition when isolating and measuring any single component. We believe that three points need to be considered with reference to the context (and background information) necessary for the reliable measurement of the components of metacognition:

1. It may be impossible—or at least theoretically naive—to study the components of metacognition in isolation.
2. “Linkage” studies (e.g., relating strategy selection and attributions) may provide the best framework for theoretical validation as well as for achieving reliable measurements.
3. There is a clear need for research in which metacognitive constructs are interrelated from a developmental perspective. The relative importance of each component in the successful integration of the entire metacognition system probably changes dramatically with age (cf. Borkowski & Thorpe, 1994).

HOW LEARNING CONTEXTS INFLUENCE THE DEVELOPMENT OF METACOGNITION

It is possible to design learning contexts that influence attributional beliefs, motivational goals, and self-efficacious beliefs as well as the efficient processing of information, eventually resulting in deep con-
ceptual understanding. We believe it is essential to consider both contextual and correlated information related to the emergence of these skills and beliefs in order to develop reliable and valid measures of metacognition.

Parents and teachers—and the learning environments they create—are pivotal to the development of an integrated metacognitive system. The beliefs that parents and teachers hold about the nature of knowledge, and about the processes related to knowledge acquisition, play powerful roles in determining the design and outcome of instructional arrangements. These experiences also have implications for both the development and measurement of metacognitive skills and beliefs.

Teachers’ Implicit Theories

Teachers’ beliefs and implicit theories about how children learn can influence their planning of daily activities and, more generally, their teaching styles. For instance, Palincsar, Stevens, and Gavelek (1989) found a complex relationship between teacher beliefs and practice in the context of teaching reading skills: Teachers who conceptualized reading as a mastery of a sequence of isolated skills tended to require children to practice strategies in a routine fashion and were content-oriented in their conceptions of reading instruction. On the other hand, teachers who were more student-oriented, devoted more time to the affective and oral language dimensions of reading instruction and, important to metacognitive development, encouraged the flexible use of strategies. The “working model” presented by Borkowski and Muthukrishna (1992) suggests that teachers entice their students to become active participants in their own learning. Because the focus of instruction is always on the child’s personally initiated learning process, the instructor needs to become adept at hypothesizing how the learner is processing information at any given moment and to adapt instructions appropriately.

In problem-centered learning contexts, as described by Muthukrishna and Borkowski (1995), Cobb et al. (1991), and Olivier, Murray, and Human (1992), teachers become committed to the belief that students need to regard mathematics, in part, as a self-constructed activity. That is, they and their classmates can learn to discover new ways to solve problems if only they make the effort to think about the subject matter and work hard in understanding problem complexity. The teacher must regard himself or herself as the critical mediator in this instructional process, designed to interr-
late key metacognitive components: skills, knowledge, beliefs, and executive processes.

The teacher makes possible maximum task involvement by prompting students to collaborate with one another in order to gain deep conceptual understanding. Teachers’ behaviors include verbalizations such as, “What do you think of what Peter just said?” “Do you agree/disagree with what Joanne said?” “Has anyone solved the problem a different way?” These verbalizations require process-oriented answers and help students to develop self-regulatory capabilities, such as monitoring, checking, and reflecting. Teachers also help students feel that they can assume personal responsibility for their own learning, by prompting them to explain and justify new solutions, resolve conflicts, and develop productive small-group relationships.

Motivation and Strategic Processing

Many researchers have argued that an understanding of motivation depends on the specification of achievement goals towards which individuals are oriented (Ames & Archer, 1987; Dweck, 1989; Dweck & Leggett, 1988; Nicholls, 1984). The achievement goal framework integrates cognitive and affective components of goal-directed behaviors. An achievement goal defines an integrated pattern of beliefs, attributions, and affect that underlies academic behavior and is represented by different ways of approaching, engaging in, and responding to achievement-related activities (Ames, 1992).

Nicholls and his colleagues (Nicholls, 1984; Nicholls, Ptaschnick, & Nolen, 1985) have identified three types of motivational orientations towards school learning. Task orientation involves a commitment to learning for its own sake: The goal is to increase understanding, to accomplish something not previously done, and to improve performance. In other words, a task orientation implies that the process of learning, including the effort involved, is an end in itself.

Task-oriented individuals strive to learn and understand, and the more they see that they have mastered a task the more competent they feel (Nicholls, 1984). With an ego orientation, the aim is to perform better than others or to establish that one’s ability is superior to another’s. In this case, learning and understanding are viewed as means to the end of establishing superiority over others. Evidence presented by Nicholls (1989) shows that these two dimensions are uncorrelated, or only slightly associated. The third motivational orientation, work avoidance, involves a desire to put forth as little
effort as possible; work avoidance is negatively related to task-orientation.

The dimensions of task orientation and ego orientation relate to students' beliefs about the causes of academic success. Thus, different achievement goals should be associated with different attributional beliefs. Nicholls, Cheung, Lauer, and Pataschnick (1989) have suggested that if students are committed to outperforming their peers, they tend to believe that superior ability or attempts to do better than others are the causes of their successes in school. Similarly, the more task oriented an individual, the more that individual believes that success in school depends on effort, interest, and attempts to understand. Whether students are oriented to one goal or the other has consequences for whether they develop a sense of efficacy and a willingness to try hard and to take on challenges, or whether they select easy tasks and give up in the face of failure. In an important paper, Ames (1992) drew attention to the need to explore how the structure of learning environments can emphasize different motivational goals and, consequently, influence how students think about themselves, their ability, their peers, and how they and their peers approach problem solving tasks.

Muthukrishna and Borkowski (1995) analyzed how a problem-centered learning environment, compatible with socio-constructivist theory (Cobb et al., 1991), may help alter existing patterns of motivational goals and beliefs as well as produce more desirable strategies. The teacher created a "sense-making atmosphere" in which mathematics was seen as a meaningful activity. The learning context fostered task orientation and the belief that success depended on attempts to make sense of the subject matter. Students were made to see that they themselves could discover ways to solve problems if only they made the effort to think about them and worked hard to understand them. Results revealed that students exposed to the problem-centered contexts rated the task-oriented goal of understanding and collaborating more highly than students in a direct explanation of strategies condition. The belief that success in mathematics derives from attempts to understand and collaborate also distinguished the problem-centered group from the direct-explanation group. Relatedly, students in the problem-centered condition reported greater use of deep-processing strategies than students in the direct-explanation group and tended to show greater evidence of strategy use on a long-term "far transfer" task. One can infer that students whose primary goal is learning for its own sake will value and use strategies that require deep processing of information. If
students are encouraged to explore and trust their intuitions, they will have a feeling of control and develop an excitement about searching for meaning and understanding—processes that promote the generalization of skills and strategies across time and settings. After all, it was in large part, to solve the problem of skill generalization that inspired the development of metacognitive theory.

Similarly, Lampert (1988) has described a research and development project in teaching mathematics that demonstrates how it is possible to foster the simultaneous construction of meaning in mathematics, task-orientation as a form of motivation, and the deep processing of information. Lampert used a lesson to demonstrate how a teacher might model a new form of social interaction that would encourage arguments among students who were learning to examine hypotheses about the mathematical structures underlying their solutions to problems. In her lessons, she presented students with problems, but did not explain how to arrive at the answers. The questions she expected of her students went beyond simply determining whether they could arrive at a correct solution. Students were expected to answer questions about the legitimacy of the strategies they had used in problem solutions. Questions were process-oriented and required students to explain and defend their strategies. In this way, Lampert stressed that strategies used for figuring out a problem were as important as the answers themselves. The role of the teacher was to engage all students in the class in forming and testing mathematical hypotheses. Lampert (1988) found that these hypotheses were embedded in the answers that students gave to a problem, and that comparing answers actually engaged students in a discussion of a wide range of hypotheses.

Collaboration in Knowledge Construction

The characteristics of learning tasks and classroom activities can have profound influences on strategy-based learning and motivational orientations, such as students' initiation about the requirements of various problems as well as the intensity and persistence with which they pursue them. Recent instructional innovations emphasize the need for students to be provided opportunities to construct knowledge and to engage in generative rather than passive learning (Brown & Campione, 1990; Bransford et al., 1990; Cognition and Technology Group at Vanderbilt, 1992; Pressley, Harris, & Marks, 1992; Schoenfeld, 1992). Learning as a social process and as a collaborative activity in pursuit of knowledge construction needs to
be stressed. In this view, children should engage in argumentation and reflection as they use and refine their existing knowledge in order to make sense of alternative points of view and to add to their knowledge base. A critical factor is that a truly collaborative learning environment demands reflection by the learners. Students are obligated to reflect on the meanings they construct and share in collaborative groups. Reflection induces an on-line awareness of one’s cognitive processes, which promotes the development of self-regulatory skills. By expressing ideas in public, by defending them in the face of questions from peers, by questioning others’ ideas, students are forced to elaborate, clarify, and reorganize their own thinking processes, contributing to the emergence of the kind of advanced cognition described by Butterfield et al. (1995).

In the “communities of learners” environment designed by Brown and Campione (1990), the aim is to produce “intelligent novices.” According to these authors, intelligent novices have “learned how to learn” rather than just to memorize facts. Intelligent novices, therefore, presumably possess a wide repertoire of strategies for gaining new knowledge. A community of learners is jointly responsible for creating knowledge as well as a learning environment that is designed to foster the development of problem solving, critical thinking, and reflective analysis.

From our vantage point, the common thread in the learning environments of Brown, Bransford, and Schoenfeld and their colleagues is that learning occurs within an active social context that promotes the emergence of executive processing skills and positive beliefs about self-efficacy. Classrooms that emphasize socially based learning differ from traditional classes in several important ways: (a) students take on more active roles in monitoring their own progress as well as that of others; (b) teachers serve as models of active learning and guide learning rather than adopting a domineering, didactic role; and (c) the content emphasis is on deep understanding rather than on acquiring a breadth of facts.

Similarly, Schoenfeld (1992) has argued for a particular agenda in order to develop classrooms that are “microcosms of mathematical sense-making” (p. 82). His problem-solving courses at the college level appear to have as their major focus the development of self-regulation, especially monitoring and control skills, as well as the development of self-directing motivational beliefs. The approach is to prompt students to monitor their solutions carefully, pursue interesting leads, and to abandon those that do not seem to result in success. Students’ ability to monitor and assess their “on-line” progress, and to act in response to these assessments, are core components of self-regulation.
The instructor’s job is to shape and structure classroom interactions (Schoenfeld, 1992). The shaping process consists of working on ideas generated by students themselves, with the teacher serving as a moderator for class discussions. A vast amount of the time is spent on collaborative efforts, either in small groups or as a whole class. Time is spent in actually doing mathematics. That is, students are engaged in the discipline, debating, conjecturing, proving, agreeing, and disagreeing. The focus is on deep levels of understanding and in enhancing positive attributional beliefs. The teacher serves as an external monitor during problem solving, encouraging discussion of behaviors considered important for the internalization of metacognitive skills, as well as a model of good executive behaviors. The hoped-for result is an increase in planning, monitoring, and active problem solving among the students.

Selecting Learning Tasks

Learning environments need to be structured so that students perform tasks that are related to interesting and coherent goals, rather than for extrinsic reasons. It is difficult to teach students to be strategic, to plan and to be cognitively alert when they are working on meaningless activities. In addition, the active use of knowledge is made clear, rather than obscured, when learning goals are personal and valuable. Brown, Collins, and Duguid (1989) have stressed the importance of “situated learning” in which knowledge is learned in the context of meaningful goals. Decontextualized forms of instruction are to be avoided. For instance, in the reading program developed by Palincsar and Brown (1984)—referred to as reciprocal teaching—comprehension monitoring strategies, such as summarization and questioning, are modeled and practiced in a context in which participants share the goal of gaining meaning from the text. The fact that students learn to apply comprehension strategies as they are being acquired is thought to be the key to the program’s success. Situated learning has a great impact on the motivational orientations students develop. Activities become more meaningful because they offer personal challenges, provide students with a sense of control over the task at hand, and create an intrinsic purpose for learning. Ames (1992) believes that if students perceive meaningful reasons for engaging in an activity, they are more likely to espouse a task-oriented goal.

Presenting learning tasks as problems to be solved rather than facts to be learned can encourage richer and more elaborated process-
ing of information, especially if this is done collaboratively. In many classrooms, however, problems are typically of a “closed” nature, and difficult for collaboration. Problems of an “open” type provide opportunities for students to share different perspectives, hypotheses, and solution paths, as well as to engage in critical analyses. Such activities influence the emergence of executive skills as well as develop positive motivational goals.

Student Perceptions of Their Learning Environments

Students need to learn that classroom activities typically require them to work hard to achieve understanding. The classroom environment must be perceived as one in which they are free to explore ideas, ask questions, and make mistakes. They should learn that it is possible, even probable, to understand what one is doing and to come to the realization that it is worthwhile and rewarding. Such an environment contributes to the emergence of short- and long-term academic and occupational possible selves (Day et al., 1992). Ames and Archer (1988) and Maehr and Midgley (1991) have argued that such visions and beliefs are likely to develop when students are involved in choice and decision making, when there are opportunities for peer interaction and cooperation, and when success is defined as much by effort and improvement as by “correctness.”

Students’ perceptions of how their responses are evaluated influence how they approach tasks and result in the development of stable orientations towards motivational goals. Brophy (1983) characterized traditional classroom learning as highly product-oriented. In contrast, there should be an emphasis placed on thinking processes: Students learn that they have a need or an obligation to process information at a deep level because they might have to explain and defend their solutions to themselves and to others.

Students must recognize that their individual ideas become of greater value when placed in a social setting. Each student is not compared with others but rather is encouraged to jointly construct meanings and solutions to problems within a social context. Peers should be seen as sources of information, rather than as threats to self-esteem. Social comparisons, when they occur, are a critical factor affecting students’ perceptions about themselves, others, and the tasks per se. Ames (1992) found that students’ self-evaluations of their ability are more negative when classroom structures emphasize winning, outperforming others, and surpassing normative standards. Social comparisons in a classroom setting can have negative conse-
quences for student interest (Deci & Ryan, 1985), pursuit of challenging tasks (Elliot & Dweck, 1988), and use of learning strategies (Ames, 1992). Relatedly, Graham and Golan (1991) found that a focus on social comparison standards can interfere with effort-based strategies that require reflective, deliberate processing. In short, collaborative-based learning seems to enhance the development of the full metacognitive system.

SUMMING-UP

It is likely that the search for domain-specific or domain-general laws about metacognition, which up to this point in two decades of research have favored the former, will likely continue to be the core issue in metacognition research and measurement. However, both intuitive appeal and scholastic relevance favor the generality side and will continue to influence the direction and style of research in this field. We suspect that the data will eventually reveal selective generality (perhaps in an executive process such as task analysis); an intricate pattern of developmentally related events necessary for achieving generality (involving consistency in metacognitively based instructions in the home and school over long periods of time); and a complex blending of specificity and generality across individuals. Not all students who have relatively similar environments will show generality and those who do may also have relative strengths and weaknesses in one and the same metacognitive component across domains. For instance, a student may be high in a variety of monitoring skills but superior in monitoring memory accuracy. Given the early stage of theory development as well as the lack of measurement sophistication that characterizes this field, it is not surprising that the search for across-tasks and across-domains generalization of metacognitive components has remained an elusive goal. Consistency in home and school instructional environments appears as the major prerequisite for developing generalized metacognitive skills and beliefs. In this sense, our discussion of the nature and quality of learning environments takes on special significance for achieving high levels of metacognitive development and stability.

Classroom environments and experiences should show each student that he or she can gain control over their own learning outcomes if they adopt self-regulatory strategies. Teachers must continually encourage students to evaluate and monitor their problem-solving initiatives. This recommendation is supported by data of Paris and Winograd (1990) who found that students will apply self-regulatory
skills if they feel that they are able to manage their own learning. Perceptions of control affect motivational, regulational, achievement processes, and outcomes as well. Finally, Grolnick and Ryan (1987) concluded that conceptual learning appears to be facilitated by contexts that minimize external controls, and at the same time focus students on the task by encouraging deep processing. Thus, students' perceptions of activities and tasks not only influence how they approach learning, but also their judgments about their ability, willingness to apply effortful strategies, and feelings of satisfaction—all of which contribute to skill-based learning.

It is not surprising that challenge, interest, and perceived control are embedded in the structure and design of problem-centered learning contexts (Muthukrishna & Borkowski, 1995). In such contexts, activities are structured as problems to be solved by all students, assisted by ample guidance, facilitation, and modeling from the teacher. Problem-centered, collaborative environments offer personal challenges and, over the long run, help students gain a sense of control, together with the emergence of task-rather than ego-orientations. Most importantly, such environments hold the potential for creating an intrinsic love of learning, housed within a mature and stable metacognitive system that yields to reliable assessment and, more importantly, gives reality to the idealized model of the “Good Information Processor” (Pressley et al., 1990).

REFERENCES


Lampert, M. (1988). The teacher’s role in re-inventing the meaning of mathematical knowledge in the classroom. In M. Behr, C. LaCampagne, & M. M. Wheler (Eds.), Proceedings of the Tenth Annual Meeting of the North American Chapter of the International Group for the
1. PROCESS-ORIENTED MODEL OF METACOGNITION

Psychology of Mathematics Education (pp. 433-480). DeKalb, IL: Northern Illinois University.


Assessing Metacognition and Self-Regulated Learning

Paul R. Pintrich

*University of Michigan, Ann Arbor*

Christopher A. Wolters

*University of Houston*

Gail P. Baxter

*Educational Testing Service*

In this chapter we provide an overview of the conceptual and methodological issues involved in developing and evaluating measures of metacognition and self-regulated learning. Our goal is to suggest a general framework for thinking about these assessments—a framework that will help generate questions and guide future research and development efforts. Broadly speaking, we see the main issue in assessing metacognition and self-regulated learning as one of construct validity. Of critical importance are the conceptual or theoretical definitions of these constructs and the adequacy of the empirical evidence offered to justify or support interpretations of test scores obtained from instruments designed to measure them.

In speaking to this issue of construct validity, we organize our chapter into four main sections. First, we review the various theoreti-
mental and conceptual models of metacognition and self-regulated learning and propose three general components of metacognition and self-regulation that will guide our discussion in subsequent sections. Second, we briefly describe a set of criteria proposed by Messick (1989) for investigating construct validity and suggest a set of guiding questions and general issues to consider in evaluating measures of metacognition and self-regulated learning. Third, we discuss in some detail several measures for assessing metacognition and self-regulated learning in light of the empirical evidence available to address issues of the construct validity of these measures. In the fourth and final section, we draw some conclusions about current measures of metacognition and self-regulated learning, suggest some directions for future research, and raise some issues that merit consideration in the development and evaluation of valid measures of metacognition.

COMPONENTS OF METACOGNITION AND SELF-REGULATED LEARNING

There is general agreement that metacognition can be divided into two general constructs termed metacognitive knowledge and metacognitive control and regulation. Some researchers have proposed that the term metacognition be reserved for the construct of metacognitive knowledge and that the term not include metacognitive control and regulation activities (Paris & Winograd, 1990). Others have proposed that monitoring and control are two different aspects of metacognition and need to be separated conceptually and functionally from each other and from metacognitive knowledge (Nelson & Narens, 1990). In this chapter, we recognize the importance of distinguishing between these different components and organize our discussion around three general components of metacognition: (a) metacognitive knowledge, (b) metacognitive judgments and monitoring, and (c) self-regulation and control of cognition. Of course, as in any model of metacognition and self-regulation, these three general components are interdependent, but for the purpose of exposition, we discuss them separately.

It should be noted that there is confusion in the literature regarding the use of the terms metacognition and self-regulated learning. Metacognition is the "older" term defined and used in the late 1970s and into the 1980s by developmental and cognitive psychologists (see Flavell, 1979). Much of the research on metacognition during this time focused on students' metacognitive knowledge about different types of memory and cognitive strategies and only later were issues
of control and regulation of cognition included (Brown, Bransford, Ferrara, & Campione, 1983). Beginning in the mid 1980s and continuing into the 1990s the construct of self-regulated learning was proposed by educational and developmental psychologists to refer to the various ways individuals monitor, control, and regulate their learning (see Schunk & Zimmerman, 1994; Zimmerman, 1986; Zimmerman & Schunk, 1989). In this research, self-regulated learning includes monitoring, controlling, and regulating cognition and monitoring, controlling, and regulating other factors that can influence learning such as motivation, volition, effort, and the self-system. Most of the models of self-regulated learning assume that the processes of monitoring, controlling, and regulating are related to, if not dependent on, metacognitive knowledge about the self and cognition (Garcia & Pintrich, 1994). As such, self-regulated learning is the more global and inclusive construct and subsumes metacognition and metacognitive knowledge. Nevertheless, we will refer to certain aspects of knowledge and monitoring as metacognitive because they are focused specifically on knowledge and monitoring of cognition. We now turn to a description of the three general components of metacognition and self-regulated learning and the various ways they have been conceptualized in the research (see Table 1).

**Metacognitive Knowledge**

Metacognitive knowledge includes students’ declarative, procedural, and conditional knowledge about cognition, cognitive strategies, and task variables that influence cognition (Alexander, Schallert, & Hare, 1991; Flavell, 1979). In some models, metacognitive knowledge is labeled as metacognitive awareness, but we believe that awareness connotes a more “on-line,” “in-the-moment,” or conscious experience and we prefer to consider that an aspect of metacognitive judgment and monitoring. We reserve the term metacognitive knowledge for knowledge about cognition and assume it is similar in many ways to other kinds of knowledge in long-term memory that individuals can have about any topic such as geography, automobiles, furniture, or mathematics. In this sense, metacognitive knowledge may be more “static” and statable than monitoring and regulation (Schraw & Moshman, 1995); that is, individuals can tell you if they know something or not, such as knowing the state capital of Nebraska or knowing the definition of words (see Tobias, this volume). In contrast, a more “on-line” measure of metacognitive monitoring would involve students’ judgments of whether they are comprehend-
ing the text or learning something about the Great Plains as they read a geography textbook.

In Flavell’s classic (1979) paper on metacognition he proposed that metacognitive knowledge included knowledge of the person, task, and strategy variables or factors that can influence cognition. In the person category he included beliefs about the self in terms of intraindividual differences (e.g., knowing that one is better at memory tasks than problem-solving tasks) as well as interindividual differences (e.g., knowing that one is better at memory tasks than a friend) and universals of cognition (e.g., knowing that one has to pay close attention to something in order to learn it). In our conceptualization of metacognition, we believe that the person variables, except for the universals of cognition, are better seen as motivational constructs (Garcia & Pintrich, 1994). They certainly represent knowledge of the self, and in that sense are metacognitive. However, because they involve the self, they are “hot” cognitions, not “cold” cognitions about task and strategy variables, and as such will not be discussed much in this chapter.

Knowledge about the task and knowledge about the strategy variables that influence cognition are the more traditional metacognitive knowledge constructs. Task variables include knowledge about how task variations can influence cognition. For example, if there is more information provided in a question or a test, then it will generally be more easily solved than when there is little information provided. Most students come to understand this general idea and it becomes part of their metacognitive knowledge about task features. Other examples include knowing that some tasks, or the goals for the task, are more or less difficult, like trying to remember the gist of a story versus remembering the story verbatim (Flavell, 1979).

Knowledge of strategy variables includes all the knowledge individuals can acquire about various procedures and strategies for cognition including memorizing, thinking, reasoning, problem solving, planning, studying, reading, writing, etc. This is the area that has seen the most research and is probably the most familiar category of metacognitive knowledge. Knowing that rehearsal can help in recalling a telephone number, or that organizational and elaboration strategies can help in the memory and comprehension of text information, are examples of strategy knowledge. In addition, metacognitive knowledge has been further broken down into declarative, procedural, and conditional metacognitive knowledge (Alexander et al., 1991; Paris, Lipson, & Wixson, 1983; Schraw & Moshman, 1995).
Table 1. Three General Components of Metacognition and Self-Regulated Learning

I. METACOGNITIVE KNOWLEDGE

A. Knowledge of cognition and cognitive strategies—knowledge about the universals of cognition
   1) Declarative knowledge of what different types of strategies are available for memory, thinking, problem-solving, etc.
   2) Procedural knowledge of how to use and enact different cognitive strategies
   3) Conditional knowledge of when and why to use different cognitive strategies
B. Knowledge of tasks and contexts and how they can influence cognition
C. Knowledge of self—comparative knowledge of intra-individual and interindividual strengths and weakness as a learner or thinker; better seen as motivational not metacognitive self-knowledge

II. METACOGNITIVE JUDGMENTS AND MONITORING

A. Task difficulty or ease of learning judgments (EOL)—making an assessment of how easy or difficult a learning task will be to perform
B. Learning and comprehension monitoring or judgments of learning (JOL)—monitoring comprehension of learning
C. Feeling of knowing (FOK)—having the experience or “awareness” of knowing something, but being unable to recall it completely
D. Confidence judgments—making a judgment of the correctness or appropriateness of the response.

III. SELF-REGULATION AND CONTROL

A. Planning activities—setting goals for learning, time use, and performance
B. Strategy selection and use—making decisions about which strategies to use for a task, or when to changing strategies while performing a task
C. Allocation of resources—control and regulation of time use, effort, pace of learning and performance
D. Volitional control—control and regulation of motivation, emotion, and environment
Declarative knowledge of cognition is the knowledge of the what of cognition and includes knowledge of the different cognitive strategies such as rehearsal or elaboration that can be used for learning. Procedural knowledge includes knowing how to perform and use the various cognitive strategies. It may not be enough to know that there are elaboration strategies like summarizing and paraphrasing, it is important to know how to use these strategies effectively. Finally, conditional knowledge includes knowing when and why to use the various cognitive strategies. For example, elaboration strategies may be appropriate in some contexts for some types of tasks (learning from text); other strategies such as rehearsal may be more appropriate for different tasks or different goals (trying to remember a telephone number). This type of conditional knowledge is important for the flexible and adaptive use of various cognitive strategies.

Metacognitive Judgments and Monitoring

Unlike the static nature of metacognitive knowledge, metacognitive judgments and monitoring are more process-related and reflect metacognitive awareness and ongoing metacognitive activities individuals may engage in as they perform a task. These activities can include four general metacognitive processes: (a) task difficulty or ease of learning judgments (EOL), (b) learning and comprehension monitoring or judgments of learning (JOL), (c) feeling of knowing (FOK), and (d) confidence judgments (see Table 1).

Individuals can make determinations of the difficulty level of the task such as how hard it will be to remember or learn the material, or in Nelson and Naren's (1990) framework what they call ease of learning judgments (EOL). These EOL judgments draw on both metacognitive knowledge of the task and metacognitive knowledge of the self in terms of past performance on the task. Further, these EOL judgments are assumed to occur in the acquisition phase of learning before the learner begins a task and therefore should be viewed separately from judgments of learning or readiness for a test (e.g., Hunter-Blanks, Ghatala, Pressley, & Levin, 1988). In the classroom context, students could make these EOL judgments as the teacher introduces a lesson or assigns a worksheet, project, or paper.

A second type of metacognitive judgment or monitoring activity involves judgments of learning and comprehension monitoring. These judgments may manifest themselves in a number of activities such as individuals becoming aware that they do not understand something they just read or heard or becoming aware that they are reading too
quickly or slowly given the text and their goals. Judgments of learning also would be made as students actively monitor their reading comprehension by asking themselves questions. Judgments of learning also could be made when students try to decide if they are ready to take a test on the material they have just read and studied. Pressley and Afflerbach (1995) provide a detailed listing of monitoring activities that individuals can engage in while reading. These types of monitoring activities are called judgments of learning (JOLs) in the Nelson and Narens (1990) metamemory framework. JOLs occur during the acquisition and retention phases in their model of memory. In each case individuals make predictions about which items on a memory task they have learned and whether they will be able to recall them in the future. In a reading comprehension task, this would involve readers, as they are in the process of reading, making some assessment of whether they will be able to recall information from the text at a later point in time (e.g., Pressley, Snyder, Levin, Murray, & Ghatala, 1987b). In the classroom context, besides reading comprehension, JOLs could involve a student making a judgment of her comprehension of a lecture as the instructor is delivering it or whether she could recall the lecture information for a test at a later point in time.

A third type of metacognitive awareness process is termed the feeling-of-knowing or FOK (Nelson & Narens, 1990; Koriat, 1993). A typical instance of FOK occurs when a person cannot recall something when called upon to do so, but knows he knows it, or at least has a strong feeling that he knows it. In colloquial terms, this experience is often called the tip-of-the-tongue phenomenon and occurs as a person is attempting to recall something. In the Nelson and Narens (1990) framework, FOKs are made after failure to recall an item and involve a determination of whether the currently unrecallable item will be recognized or recalled by the individual at a later point in time. Koriat (1993) points out that there is evidence that FOK judgments are better than chance predictors of future recall performance, albeit not a perfect correlate. In a reading comprehension task, FOKs would involve the awareness of reading something in the past and having some understanding of it, but not being able to recall it on demand. FOKs in the classroom context could involve having some recall of the teacher lecturing on the material or the class discussing it, but not being able to recall it on the exam.

A fourth type of metacognitive judgment concerns the confidence an individual has in their retrieved answer on a memory task, a reading comprehension task, or even on a classroom exam. This
confidence judgment is assumed to come after some retrieval of information and some output response or behavior has been enacted (Nelson & Narens, 1990). For example, students might be given a text to read, asked to answer some questions about it, and then asked to judge the confidence they have in their answers (Pressley, Ghatala, Woloshyn, & Pirie, 1990). Another type of confidence judgment has been used in error detection studies. Students are given a text to read that has errors in it and they are asked to find contradictions or errors in the text. After they have finished reading the text and reporting on the errors they found, students are asked to rate their comprehension of the text and rate their performance in detecting the errors (Baker, 1989b). These judgments of comprehension and error detection performance are assumed to reflect some metacognitive awareness about the correctness of performance and the calibration of these confidence judgments to actual performance is an important aspect of metacognitive judgment and monitoring.

Self-Regulation and Control

The types of activities that individuals engage in to adapt and change their cognition or behavior are known collectively as self-regulation and control. In this sense, this component is more of a process, ongoing activity, like metacognitive judgments and monitoring, than a static entity like metacognitive knowledge. In most models of metacognition and self-regulated learning, control and regulation activities are assumed to be dependent on, or at least strongly related to, metacognitive monitoring activities, although metacognitive control and monitoring are conceived as separate processes (Nelson & Narens, 1990; Zimmerman, 1989, 1994). In this chapter we focus on measures of control and regulation of cognition that could be more narrowly labeled metacognitive control and self-regulation. Other aspects of self-regulated learning including motivation, effort, volition, goals, and the self-system, can be "controlled" and therefore are included in our framework of self-regulated learning (see Table 1). However, because control and regulation of these components have not been studied as much as control and regulation of cognition, they are not discussed in as much detail in the third section of this chapter on construct validity of the instruments to measure metacognition and self-regulated learning.

In the Pressley and Afflerbach (1995) model of constructively responsive reading, monitoring activities include monitoring of comprehension as well as a variety of decisions to change reading strate-
gies and behavior such as varying the speed of reading, rereading, or taking notes on reading material. This model is based on data from in-depth verbal protocol analyses of reading behavior where it is clear that monitoring and regulating activities often occur at the same time. Likewise in our self-report data on metacognition and self-regulation, it has not been possible to separate empirically cognitive monitoring from control and regulation of cognition (Pintrich & De Groot, 1990; Pintrich, Smith, Garcia, & McKeachie, 1993). Despite the empirical difficulties demonstrated by these studies, conceptually it is possible to distinguish between monitoring activities that involve assessing comprehension, learning, or performance, and regulating activities that involve changing cognition or behavior to bring them in line with personal goals or task demands. Further, there are a number of different activities that can be considered part of the various control and regulation processes. We organize our conceptual discussion around the four general categories of planning, strategy selection and use, resource allocation, and volitional control (see Table 1).

Planning is an important aspect of regulating cognition and behavior and involves the setting of goals that can be used to guide cognition in general and monitoring in particular (Pressley & Afflerbach, 1995; Schunk, 1994; Zimmerman, 1989; Zimmerman & Martinez-Pons, 1986, 1988). The goal acts as a criterion against which to assess and monitor cognition, just as the temperature setting of a thermostat guides the operation of the thermostat and heating/cooling system. For example, if one student has a goal of mastering the text material as opposed to another student who just wants to complete the reading assignment, then the first student will monitor and regulate her reading cognition in a way that can lead to deep understanding (e.g., use self-questioning or reread parts that are not understood). In contrast, the second student may just proceed to read through the material and, when at the end of the selection, be satisfied that the goal of completing the reading has been reached. Of course, planning is most often assumed to occur before starting a task, but goal-setting can actually occur at any point during performance. Learners may begin a task by setting specific goals for learning, goals for time use, and goals for eventual performance, but all of these can be adjusted and changed at any time during task performance.

One of the central aspects of the control and regulation of cognition is the actual selection and use of various cognitive strategies for memory, learning, reasoning, problem solving, etc. Numerous studies have shown that the selection of appropriate cognitive strategies can have a positive influence on learning and performance. These
cognitive strategies range from the simple memory strategies very young children through adults use to help them remember (Schneider & Pressley, 1989) to sophisticated strategies that individuals have for reading (Pressley & Afflerbach, 1995), mathematics (Schoenfeld, 1992), writing (Bereiter & Scardamalia, 1987), problem solving, and reasoning (see Baron, 1994; Nisbett, 1993). Although the use of various strategies is probably deemed more "cognitive" than metacognitive, the decision to use them is an aspect of metacognitive control and regulation as is the decision to stop using them or to switch from one strategy type to another.

The third aspect of self-regulation and control that we include in our framework is the allocation of resources such as time, overall effort, and pace of learning. These resources may not be strictly cognitive because they do not involve specific cognitive strategies, but the control and regulation of these resources can be an important aspect of self-regulated learning (Nelson & Narens, 1990; Pintrich, Smith et al., 1993; Zimmerman, 1989, 1994). Obviously, a greater amount of time spent studying a list of words to be memorized or a set of text materials for an exam should result in improved learning and performance. Moreover, the amount of overall effort put into a task can reflect overall time use, the intensity of study including the use of more appropriate cognitive strategies, or more attention and concentration on the task without the use of better strategies. Finally, the pace of learning, how fast individuals perform the various subtasks of the overall task, is an important feature that self-regulated learners can control.

A fourth category of self-regulation and control is what we have called volitional control. Although some theorists have termed all of the metacognitive control and regulation activities as volitional control (cf. Corno, 1993; Kuhl, 1985, 1992), we reserve this term for the control of emotion, motivation, and the general environment. As learners engage in tasks, their cognition, emotions, and motivational beliefs are activated. Consequently, the learners’ ability to control and regulate their emotions can play an important part in their learning (Pressley & Afflerbach, 1995). In the same manner, motivational beliefs can have a dramatic influence on cognition, learning, and performance (Pintrich, Marx, & Boyle, 1993; Pintrich & Schrauben, 1992) and attempts to regulate or control motivation could result in improved learning. Both Corno (1993) and Kuhl (1985; 1992) have suggested that individuals’ ability to control their environment (e.g., arrange for quiet space for studying away from distractions) is an important aspect of self-regulation. Although the control of motiva-
tion and emotion are important aspects of self-regulated learning, we do not discuss in much detail the various instruments to assess them in this chapter because of our focus on the cognitive components of self-regulated learning, not the motivational components.

Taken together, planning, strategy selection, resource allocation, and volitional control comprise four important aspects of self-regulation and control. In combination with metacognitive judgments and monitoring, they make up the "on-line" process-oriented aspects of metacognition and self-regulated learning. The "static" component of metacognition, metacognitive knowledge, once activated in a situation, is an important resource that is drawn upon by learners as they monitor and control their own learning. In proposing this three component model of metacognitive knowledge, monitoring, and self-regulation and control, and their corresponding subcomponents, we lay a conceptual framework for examining the empirical evidence for the construct validity of our measures. We turn now to a discussion of construct validity.

CRITERIA FOR EVALUATING CONSTRUCT VALIDITY AND RELEVANCE/UTILITY OF A MEASURE

One of the fundamental issues in evaluating assessment instruments purporting to measure metacognition and self-regulated learning is that of construct validity. Historically, construct validity was conceived as one of three essential aspects of validity termed construct, criterion (predictive and concurrent), and content validity. Each aspect was defined to some extent with respect to the purpose of the measure. Content validity was of primary importance for achievement tests where issues of the overlap between test items and a subject-matter domain were addressed by professional judgment. Tests designed to predict future performance (e.g., success in college) or tests designed to replace an existing measure, relied on criterion-related validity evidence typically in the form of data from correlations or regressions where the test score (e.g., SAT score) was related to the "criterion" (e.g., success in college as measured by undergraduate GPA). Construct validity in the form of correlational, experimental, or other forms of data analysis argued for the presence of latent or unobservable traits such as anxiety or intelligence.

Recent conceptions reject this traditional three-pronged approach in favor of a "unified" validity theory with construct validity as the overarching issue and all other "types" of validity subsumed under it (Cronbach, 1989; Cronbach & Meehl, 1955; Messick, 1989; Shepard,
In his comprehensive treatise of validity, Messick (1989) restated the centrality of construct validity and drew attention to its relations to the value and consequences of test interpretation and use. In an effort to clarify these relations, he proposed a four quadrant model of validity that crosses the nature of the empirical evidence on the test and the potential consequences of the test data with how the test is interpreted and used (see Figure 1).

**Figure 1. Messick’s (1989) Conceptualization of Construct Validity.**

In Cell 1, Messick (1989) considers a number of specific types of evidence that can be offered to support test score interpretations. Collectively, he terms these different types of evidence as construct validity. Essentially, construct validity involves a determination of how well the instrument produces scores that avoid two basic measurement problems: (a) construct underrepresentation or not measuring all relevant aspects of the construct and (b) construct irrelevant variance or measuring other constructs, not just the target construct. Moving across the row, Cell 2 considers evidence required for test score use and includes not only construct validity but relevance and utility of the scores as well.
In the bottom two cells in Table 2, Messick has placed concerns about the consequences of the uses of test scores. In terms of test score interpretation, Cell 3, labeled value implications, concerns how the construct is defined theoretically and conceptually and the ways in which this theoretical framework reflects underlying societal values or ideologies. In Cell 4, Messick considers the social benefits and costs of using the test scores. For example, intelligence tests and achievement tests are often used to classify children for special services or for selection and placement into different academic tracks and each of these decisions has a number of social costs associated with it. Given that most measures of metacognition or self-regulated learning are not used in this manner, we focus our comments on the first row in Figure 1 and consider issues of construct validity and relevance/utility of measures of metacognition and self-regulated learning. In what follows, we describe each more fully, noting where appropriate similarities and differences between Messick’s formulation and the writings of others (e.g., Linn, Baker, & Dunbar, 1991; Shepard, 1993).

Messick (1989) proposes five general components of construct validity that merit consideration under the heading evidential basis for test score interpretation. These are content, substantive, structural, external, and generality of meaning. These five components are interdependent and although it is difficult to draw sharp distinctions or boundaries between them, we discuss them separately.

Content Component. This component is concerned with the relevance and representativeness of the content coverage of the assessment tool in relation to the domain of interest (Messick, 1989). The basic guiding question is: Are the items on the test representative of the domain? In achievement testing this concerns how well the content of the test reflects the content of the domain. Linn et al. (1991) suggest that there are three important aspects that should be considered in examining content validity: (a) domain specification, (b) relevance and meaningfulness of the tasks, and (c) representativeness of the content.

Domain specification concerns assumptions that the test has important content on it. For example, would the content on the test be considered important by most individuals in the field? Relevance and meaningfulness of the tasks concerns the assumption that instructional relevance stems from students being asked to do tasks that are as meaningful, relevant, and authentic as possible. Although this aspect may not be crucial for paper-and-pencil tests, this is especially important for performance assessments where students are often asked to engage in extended tasks or solve complex problems that demand
sustained thinking and reasoning. Finally, the issue of representativeness of content concerns the comprehensiveness of the content coverage of the test relative to the subject-matter domain. For example, in constructing a science achievement test, one should consider the degree of overlap between the content included on the test and the various domains and important concepts students have learned in science.

In terms of metacognition and self-regulated learning, content coverage and representativeness are important issues when considering measures of metacognitive knowledge. Metacognitive knowledge can include knowledge of strategies and conditional knowledge of when and why to use these strategies. Given that there are probably many different strategies for learning in the domains of literacy, mathematics, science, and social studies, there will be different domains of metacognitive knowledge. Accordingly, assessments of metacognitive knowledge must be examined in terms of their content coverage for a particular domain. The issue of representativeness may not be as important if only one domain is under consideration (e.g., reading) unless the measure is assumed to be a general measure of metacognitive knowledge but only assesses metacognition for reading words. Metacognitive monitoring and self-regulation and control are usually assumed to be general, content-free processes. Consequently, issues of content representativeness are of less importance for these assessments than for measures of metacognitive knowledge.

Substantive Component. Whereas the content component concerns the relation between the test items and the larger content domain from which the items are sampled, the substantive component refers to the internal relations between the data generated from the items and the construct theory. The basic guiding question is: Are the response patterns on the test consistent with the theory of the construct? In particular, Messick (1989) notes that items on the test as well as individuals' responses to these items should exemplify the construct being measured and not other constructs. Further, items that ostensibly measure a different construct should not be related to the test items of the targeted construct. In achievement testing, for example, the items on a science achievement test should reflect to a large extent acquired knowledge in science, not general intelligence or general reading ability. The same logic applies to measures of metacognition. For example, measures of metacognitive monitoring and awareness should assess monitoring and not other constructs such as verbal ability, prior knowledge, or general intelligence (Pressley & Ghatala, 1990). Accordingly, a measure of metacognitive monitoring that is dependent on the learner’s verbal fluency and ability to
articulate their thinking and awareness may be introducing construct-irrelevant variance into the measure.

Structural Component. This component of construct validity concerns the relations between the scoring system used and the theoretical model of the construct. The guiding question is: Do the scores on the test and the scoring system reflect the complexities of the construct as expressed in the theoretical model? Generally, the relations between items on a test—how they are scored and then summed, combined, or kept separate—should reflect the same relations as those expressed in the theoretical model of the construct. A single total score from a test implies that the construct is unitary; a number of different subscales or scores implies a differentiated construct; a combination of one total score with several subscales implies a hierarchical model of the construct under consideration.

In achievement testing, separate subscores for different aspects of mathematics such as geometry, algebra, and trigonometry implies that there are different domains of expertise and separate scores are necessary to capture this complexity. On the other hand, one total score assumes that there is an overall general mathematical expertise construct. Of course, it is possible to have a conceptual model that underscores the utility of having scores that express domain-specificity as well as general expertise in mathematics. In terms of measures of self-regulation and control, if a model proposes a general construct as well as the four subprocesses of planning, strategy selection, resource allocation, and volitional control, then there may be a rationale for having one general self-regulation score and four subscores corresponding to the four subprocesses.

Another issue that is subsumed under the dimension of structural validity concerns the interpretation of scores in terms of normative versus ipsative models and criterion-referenced versus norm-referenced models (Messick, 1989). The normative-ipsative distinction represents the measurement version of the nomothetic-idiographic distinction made in psychology and education. Normative models are concerned with how individuals differ with respect to some variable or construct, allowing for comparisons between individuals. In contrast, ipsative models order scores on some set of attributes or variables for each individual allowing for intraindividual comparisons of the relative strengths or weaknesses across attributes (Messick, 1989). In a similar fashion, norm-referenced scoring models highlight the distribution of scores and allow for comparisons between individuals on the construct. For example, intelligence as a theoretical construct is usually conceptualized in a normative fashion and IQ
scores are usually scaled to facilitate interpretation of an individual’s score relative to the population distribution of IQ scores. In contrast, criterion-referenced scoring models allow for the comparison of an individual’s score to some standard and individuals are judged in relation to that standard, not with respect to how others performed (e.g., driving test).

In terms of measures of metacognition and self-regulated learning, the Learning and Study Skills Inventory or LASSI (Weinstein, Schulte, & Palmer, 1987; Weinstein, Zimmerman, & Palmer, 1988) uses a norm-referenced system so that students’ responses can be compared against a normative sample. In this case, there is an underlying theoretical assumption that students’ scores are somewhat general and stable across situations, allowing for normative comparisons. Other measures such as the Motivated Strategies for Learning Questionnaire or MSLQ (Pintrich & De Groot, 1990; Pintrich, Smith, et al., 1993) or the Self-Regulated Learning Interview Schedule or SRLIS (Zimmerman & Martinez-Pons, 1986, 1988) do not use norms. These measures reflect a theoretical assumption that students’ responses may vary as a function of the task, situation, course, or school context, thereby rendering normative comparison groups less useful.

**External Component.** This component of construct validity asks the basic guiding question: “What is the nature of the relations between and among various measures and the construct of interest?” Evidence may come from correlational studies of the pattern of relations among measures that purport to measure similar or different constructs with similar or different methods (cf. Campbell & Fiske, 1959). In addition to these multitrait multimethod studies, evidence may come from an examination of the actual and theoretically predicted relations between measures of different constructs. Also known as nomological validity, the issue is one of fit between theory and observed relations between tests scores and other measures of the construct.

In achievement testing, this might involve the collection of data to demonstrate how well the test scores relate to the grades students receive from teachers or how long they have been in school or how many courses they have taken in that domain and their performance in those courses. At the same time, scores on the achievement test might be compared to other general ability measures (intelligence) and the data should show moderate positive relations given that achievement and intelligence are usually conceptualized as separate constructs, albeit our theories predict they will be positively related. In terms of metacognition, if metacognitive monitoring is assumed to
be an important component of skilled reading, then measures of metacognitive monitoring should be positively related to other measures of reading performance such as reading achievement test scores, teachers rating or grades for reading, and measures of reading comprehension.

Generality of Construct Meaning. The guiding question for this component asks how generalizable the scores are across different populations (e.g., males and females; different ethnic groups), domains (e.g., mathematics and reading), tasks, settings, and time. In terms of population generalizability, the issue is whether the assessment data from different groups of students can be scored and interpreted in the same way. Differential performance of various populations of students (differing by gender, ethnicity, etc.) has always been a concern in achievement testing. Recently, generalizability has been cited as a primary limiting factor in the use of performance assessments—as a replacement or supplement to 40-item multiple-choice tests common in educational measurement (e.g., Linn et al., 1991; Shavelson, Baxter, & Gao, 1993). For measures of metacognition and self-regulated learning there is a great deal of evidence to suggest that metacognition and self-regulated learning change with age, both in level and quality, and assessment instruments must take this into consideration. Further, consideration must be given to the consistency of measures across groups varying in gender, ethnicity/culture, or socioeconomic status (SES).

Also included under this component are issues of how the assessment data generalize across domains and tasks (see Linn et al., 1991). For example, in terms of domain specificity, does a science performance assessment score for solving a circuits problem generalize to other aspects of science performance in earth science or biology? In terms of task specificity, does the score on a paper-and-pencil measure of students’ knowledge of circuits correspond to their performance on a hands-on performance assessment on the same content (Baxter & Shavelson, 1994; Gao, Shavelson, Brennan, & Baxter, 1996). This issue of domain and task generalizability is one of the major unresolved issues in our theories of metacognition and self-regulated learning and consequently as well with our assessment procedures. Empirical studies are mute with respect to the relation between a person’s metacognitive monitoring score on a reading comprehension task and her metacognitive monitoring score on a mathematics problem-solving task. High correlations may support a domain-general theory whereas low correlations may support a domain-specific interpretation. Further, inconsistent results with different methods of
assessing metacognition within a domain such as think-alouds, cloze procedures, or multiple-choice questionnaires may arise from construct relevant or irrelevant variance. Finally, in terms of temporal generalizability, for the same tasks in the same domain, there should be some consistency across time in individuals’ performance, at least within a restricted range of time where development or learning opportunities are minimal (cf. Ruiz-Primo, Baxter, & Shavelson, 1993).

Issues of domain, task, and time generalizability are more or less important depending on the theoretical stance adopted regarding the situational nature of metacognition and self-regulated learning. If the conceptual model assumes that all cognition and behavior are always and mainly situational, then there is no expectation that there will be much consistency across domains, tasks, or time. Consequently, variations in scores across these contexts are viewed positively, or at least as non-problematic. At the other end of the continuum, if the conceptual model assumes that metacognition is a stable personal “trait” of the individual, then there should be a fairly high level of consistency across contexts and deviations from consistency are viewed as problematic for the theory and the assessment instrument.

Relevance and Utility Concerns in Test Use. Besides the more technical aspects of construct validity, Messick (1989) suggests that the meaning, relevance, and utility of a measure must be considered once the test is prepared for actual use. Linn et al. (1991) suggest that for performance assessments, utility concerns the purpose of the measure, issues of cost and efficiency, and ease of use. In terms of purpose, a distinction can be made between measures of metacognition and self-regulated learning designed primarily for research purposes (i.e., to understand and analyze the various components of metacognition and self-regulation) and those used to improve practice (i.e., to gauge general levels of student metacognition and self-regulation in the classroom or for diagnostic purposes). Some methods, such as think-aloud protocols, may be more easily used in the laboratory or controlled settings such as one-to-one interviews that take place in schools, but outside the classroom. Other methods, such as questionnaires or self-reports, can be used in whole group settings such as classrooms without too much disruption to established routines. Regardless of the purpose for which the method was designed, each varies in terms of ease of use and cost. Self-report questionnaires are relatively easy and inexpensive to administer and score in terms of labor and time; think-alouds and interviews require extended periods of time and trained personnel for both administration and scoring.
A number of different instruments for assessing students' metacognition have been developed. In this section we discuss several of these instruments in light of our conceptual framework for metacognition and self-regulated learning and Messick's proposed framework for assessing construct validity. In particular, we focus on construct validity and issues of relevance and utility described above. Consistent with the three-component model of metacognition and self-regulated learning described in the first section of this chapter, we consider first measures of metacognitive knowledge, then measures of metacognitive monitoring, and finally measures of control and self-regulation. For each type of measure, we report relevant empirical studies that bear on issues of construct validity. Our purpose is to illustrate the problems and the accomplishments associated with establishing evidence of construct validity for measures of metacognition and self-regulated learning. In doing so, we set aside a comprehensive review of all measures and corresponding empirical work in favor of attention to selected measures that exemplify key issues in evaluating assessments of metacognition and self-regulated learning.

**Measures of Metacognitive Knowledge**

Because the knowledge component of metacognition is much like other static knowledge stored in long-term memory, measures to assess it can look quite similar to standard tests of subject-matter knowledge. For example, the Index of Reading Awareness (IRA), developed by Paris and his colleagues, is a multiple-choice questionnaire designed to measure metacognitive knowledge in the domain of reading comprehension (Jacobs & Paris, 1987; Paris & Jacobs, 1984; Paris & Myers, 1981). The 20-item instrument, designed for use with elementary school children, consists of 5 items in each of four sections: (a) self-knowledge and task knowledge about reading (evaluation), (b) knowledge of planning and skimming (planning), (c) knowledge about changing and adjusting reading behaviors (regulation), and (d) knowledge of when one might use different reading strategies (conditional knowledge).

In taking the IRA, students are posed a question and asked to choose one of the three possible responses. An example of a regulation item is: “What do you do if you come to a word and don’t know what it means?” For all 20 items, each of the three choices are
assigned scores of 0 for an inappropriate answer, 1 for a partially appropriate answer, and 2 for the best or most strategic answer (Jacobs & Paris, 1987). For the question above, the responses were categorized as: Best response (2 points)-Use the words around it to figure it out; Partial credit (1 point)-Ask someone else., and No credit (0 points)-Go on to the next word. Scores for each of the 20 items are then summed and higher scores are interpreted as reflective of more metacognitive knowledge.

The Metacognitive Assessment Inventory (MAI), developed by Schraw and Dennison (1994), attempts to tap into metacognitive knowledge in a somewhat different manner than the IRA. The MAI presents college students with 52 different items grouped into two scales termed general metacognitive knowledge and regulation of cognition. As an example, one knowledge item on the MAI states, “I have a specific purpose for each strategy that I use.” Students are asked to indicate how true or false each statement is for them on a 100 mm line where 0 indicates not true at all and 100 mm indicates very true for me. Scores are computed by averaging the lengths of the line for items corresponding to each scale.

The IRA is similar to a multiple-choice test, whereas the MAI is similar to a traditional self-report instrument. Taken together, empirical studies of these two instruments help to illustrate some of the issues that must be addressed when considering the construct validity of measures for assessing metacognitive knowledge. In what follows, we review research studies for each of these instruments using Messick’s framework described in the previous section as an organizational guide.

Content Component. The IRA and the MAI provide a good contrast between a domain-specific and a more general measure of metacognition. Establishing evidence for the content validity of these instruments involves determining how well each covers the intended domain. The IRA is designed to assess metacognitive knowledge in the area of reading comprehension. Pressley and Afflerbach (1995) list over 150 different activities that skilled readers engage in as they read. In assessing metacognition in this context, how large a sample of items is needed to tap adequately the important components of these 150 activities? Are 20 items sufficient? Generalizability studies would provide important information on the extent to which the items on the test generalize to the larger domain of metacognitive knowledge (cf. Shavelson, Gao, & Baxter, 1995).

In contrast to the IRA, items on the MAI are not tied to any specific domain such as reading, but instead focus on more general
learning situations and hence more general metacognitive knowledge. The content validity of this instrument depends on how well general metacognitive knowledge is sampled. The MAI includes 17 items aimed at assessing students' declarative, procedural, and conditional knowledge in addition to other items that measure aspects of metacognitive monitoring and control. Again, as for the IRA, questions as to the adequacy with which the 17 knowledge items adequately sample the domain of general metacognitive knowledge have not been answered empirically.

**Substantial Component.** Substantial validity concerns the match between the data generated by the items on the test and the construct theory. In terms of the IRA, the conceptual model predicts four subcomponents of metacognitive knowledge in reading: evaluation, planning, regulation, and conditional knowledge. Although Jacobs and Paris (1987) did not report factor analysis results or alphas for the four subscales of the IRA, a study by McLain, Gridley, and McIntosh (1991) of third, fourth, and fifth graders reported extremely low alphas (between .15 and .32) for the four subscales of the IRA. These results suggest that the four subscales of the IRA, although theoretically important, lack empirical support as four independent subcomponents.

Schraw and Dennison (1994) found a similar pattern of results with the MAI. Although their conceptual model predicted eight subcomponents including three subscales for knowledge (declarative, procedural, and conditional), results of factor analyses in two different studies of college students supported the use of only one knowledge scale and one regulation scale. These two scales had high internal consistency producing alphas of .88 and .91, for knowledge and regulation, respectively. This theory-data mismatch is a continuing problem in the field. There seem to be more factors or components predicted by theory than supported by the data generated from the empirical studies of the instruments.

This mismatch between theory and empirical data can be conceived as a problem in "grain size" or resolution power as suggested by Howard-Rose and Winne (1993). That is, our theoretical models have proposed relatively fine distinctions, or small grain-size components of metacognition. However, our instruments may not be powerful or precise enough to bring these smaller grain-size components into resolution. It remains an issue for future research and development to determine if we need to develop more powerful "microscopes" to observe these smaller grain-sized units or whether we need to modify our theoretical models to reflect the functional
nature of the fairly molar components of metacognition and self-regulated learning that seem to emerge from our data.

**Structural Component.** An important aspect of structural validity is the way in which an instrument is scored, and in turn how scores are combined. For the IRA, students are given 0, 1, or 2 points for each item based on the appropriateness of the response they select from three possible options. Points are summed to create a subscale score for each of the evaluation, planning, regulation, and conditional knowledge scales. The combination of these four scales results in a total score for the entire instrument. At the question level the scoring system is basically ordinal, the 2-point response is judged to be superior to the 1-point response, which is considered superior to the 0-point response. Nevertheless, the types of analyses carried out assume an interval scale. In the absence of a good theoretical model for differentiating the quality and quantity of metacognitive knowledge, it is difficult to defend using interval or ordinal scaling metrics. Most of our theoretical models simply assume that more metacognitive knowledge is better, hence, the summative scoring on both the MAI and IRA. However, it may be more adaptive to have metacognitive knowledge that is situation- or task-specific, but we have not developed tasks and scoring rubrics or metrics that can capture these types of conditional relations between metacognitive knowledge and different tasks. Needless to say, this is an area that is ripe for further research and development activity.

**External Component.** External validity is a reflection of how well performance on one measure is related to other measures of the same or different constructs. Paris and Jacobs (1984) and Schraw and Dennison (1994) attempted to provide some evidence that speaks to the external validity of their respective instruments by examining the relation between students' metacognitive knowledge and their standardized achievement test scores. Schraw and Dennison (1994) used portions of the Nelson-Denny vocabulary and reading comprehension tests with the MAI, whereas Paris and Jacobs (1984) used the Gates-McGinitie test of reading achievement and McLain et al. (1991) used the Woodcock test of reading in their studies of the IRA. In all cases, the authors found a positive, but modest relation between their respective measure of metacognitive knowledge and students' standardized achievement scores, with correlations ranging from .20 to .35. Correlations with standardized achievement tests provide some evidence to support the external validity of these instruments, but they should not be the sole criterion used in this regard because these rather global and stable measures may not be sensitive to variations
in metacognitive knowledge (Jacobs & Paris, 1987). There are other measures of performance and metacognition that might be expected to show positive relations. For example, Paris and Jacobs (1984) examined how students’ scores on the IRA were related to their performance on both cloze and error-detection tasks, tasks that require more explicit metacognitive skills. These analyses showed that scores on the IRA were positively related to these measures of reading comprehension and were of the same magnitude as correlations with standardized achievement tests.

Another type of evidence that bears on this issue of external validity is the comparison of pre-existing groups or groups that are assigned to treatments that are thought to vary on the construct. For example, Paris and his colleagues found that the IRA distinguished between students who were classified as good and poor readers a priori (e.g., Cross & Paris, 1988; Paris & Jacobs, 1984; Paris & Oka, 1986). Good readers were much more likely to have higher IRA scores. The IRA also distinguished between those students who received a specific curriculum designed to increase metacognitive knowledge and use of cognitive strategies and students who did not receive this program, with as expected, those in the metacognitive curriculum having higher IRA scores. As Messick (1989) notes, experimental studies of different types of students or students in different educational programs can add greatly to the evidence for the construct validity of the measure. Experimental studies are not used as frequently as correlational studies, but given the relative yield, experimental studies should be used more often in construct validity research on metacognition and self-regulated learning.

Generality of Construct Meaning. Assessments of metacognitive knowledge have, for the most part, been designed for use with a particular age group within a particular domain. The IRA, for example, was designed for use with elementary school children who are beginning to read. Although studies conducted with the IRA have included large numbers of subjects, the studies have not included children from different racial/ethnic or ability backgrounds. Examining a more diverse subject population would provide insight into the generalizability of the construct across different populations of students. For example, Swanson (1993) examined the metacognitive knowledge of students classified as learning disabled, normative, or gifted. Differences in the degree to which metacognitive knowledge and problem-solving abilities were intercorrelated within these groups suggest that the meaning of the construct, in terms of its relation to other constructs, varies in different populations of students.
The MAI also has seen limited use in nontraditional populations, in part perhaps because the instrument is new. Nevertheless, similar points about generality of construct meaning can be made. Recall that the IRA was restricted to early elementary school students. The MAI has been used primarily at the other end of the educational spectrum, college students. College students are a select group of late adolescents and generalizations to this age group in the general population are questionable. The use of undergraduate college students at a single university demonstrates a recurring generalizability issue in much psychological research. Examining metacognitive knowledge in groups or ages that extend the usual boundaries of samples of white, middle class students (cf. Graham, 1992) will provide evidence to support the construct generality of the various instruments.

Relevance and Utility Concerns for Test Use. The IRA and the MAI can be readily used in a classroom or group setting because they are easy to administer and most students are quite familiar with the response formats on each of these measures. Relative to other formats for assessing metacognition or self-regulation such as think-alouds or interviews (to be discussed below), self-report questionnaires are easy for teachers and students to use and can provide information about a large number of students in a practical and efficient manner.

Summary. According to a number of researchers, metacognitive knowledge is similar to other knowledge in long-term memory and can be accessed by the individual when properly cued (Alexander et al., 1991; Flavell, 1979). Thus, self-report instruments such as the IRA and the MAI seem appropriate for obtaining this information. The ease and efficiency with which these measures can be administered and scored facilitate their use in educational and research settings. At the same time, there remain significant questions and concerns about the construct validity of these measures. First, the content representation of the items on these two instruments may not be adequate given the rather large domain of metacognitive knowledge they purport to measure. Second, there is a continuing mismatch between the theoretical models of metacognitive knowledge that propose multiple dimensions or subcomponents and the empirical data that often yields one general factor or scale of metacognitive knowledge. Third, there is a need for theoretical work on how best to conceptualize a metric for quantifying metacognitive knowledge, followed by the concomitant psychometric research to validate new scaling procedures. Fourth, although there is more research on the relations with standardized achievement tests and comparisons of different groups of students for measures of metacognitive knowledge than other
components, studies that include other constructs as external criteria (e.g., intelligence) would be useful. Finally, there is a great need for studies that examine the generalizability of these measures for groups of students that differ on age or ethnicity or educational category such as “at-risk” students.

Measures of Metacognitive Judgments and Monitoring

The awareness or monitoring aspect of metacognition reflects an “on-line” process that includes students’ current thinking, awareness, consciousness, or monitoring of their cognitive operations just before, during, or just after completion of a task. There have been a number of different methods used to assess this aspect of metacognition including self-report of monitoring-based judgments (see Baker, 1989b; Nelson & Narens, 1990; Tobias, this volume; Tobias & Everson, 1995), error-detection studies, interviews, and think-aloud protocols (Pressley & Afleerbach, 1995). We first provide general descriptions of these different measures and then an analysis of the empirical evidence for construct validity.

Self-report. Nelson, Narens, and their colleagues carried out a series of studies using self-report judgments to measure student monitoring (Leonesio & Nelson, 1990; Nelson, 1996; Nelson, Gerler, & Narens, 1984; Nelson & Narens, 1990). Generally, students are presented with some information to be retrieved later (e.g., a list of words, a paired associates recall task). Before they actually perform the memory task they are asked to rank or rate how easy the information will be to learn (an ease-of-learning judgment or EOL). Then, these subjects are given a number of learning/study trials where they learn the list to criterion. After the learning trials, students are asked to rank or rate their level of learning, or to make a judgment of their level of learning (a judgment-of-learning or JOL). Students are then given a retention test and are told which items they did not recall. After receiving this feedback on their performance, students are asked to rank or rate which of those unrecollected items they think they may know. These judgments are called feeling of knowing (FOK) judgments (Leonesio & Nelson, 1990; Nelson et al., 1984; Nelson & Narens, 1990). Students’ confidence in their performance is usually assessed after a performance; students are asked to make some rating or assessment of how well they did on the task. Taking actual performance as the standard, the accuracy of these judgments is considered an indicator of students’ monitoring ability. Thus, students who felt they knew something and did, as well as students who felt they did
not know something and did not, are both considered good monitors of their performance. The assumption is that the ability to make accurate judgments of what one knows and what one does not know is an important aspect of metacognitive monitoring.

Using a similar judgment method with a different set of tasks, Pressley and his associates asked students how well they performed just before, during, or just after completing memory or reading tasks (Hunter-Blanks, et al., 1988; Pressley et al., 1990; Pressley, Levin, Ghatala, & Ahmad, 1987a; Pressley et al., 1987b). For example, Pressley et al. (1987b) reported three experiments in which undergraduate students read short passages from an introductory psychology textbook and predicted their level of performance on either multiple-choice or fill-in-the-blank questions. In line with the constructs from the Nelson and Narens (1990) framework, these types of studies assess students’ judgments of learning (JOL) because they ask for an assessment of current learning.

In contrast to the JOL measures of current learning, Tobias and his colleagues asked students to make judgments of prior learning—what they already know—about word knowledge or mathematics problem solving (see Tobias, this volume; Tobias & Everson, 1995; Tobias, Hartman, Everson, & Gourey, 1991). In their studies of word knowledge, students are shown a list of words and then asked to check one of two boxes indicating whether they know the definition of the word or they do not know the definition. Similarly for the mathematics problem-solving task, students are shown a set of mathematics problems and asked to check one of two boxes indicating whether they can solve the problem or they cannot solve the problem. Students are asked to go through these estimates quickly; judging 30 mathematics problems in 6 minutes or about 12 seconds per problem.

*Error detection.* In work directed by Baker, an error detection methodology was used to assess metacognitive monitoring (Baker 1979, 1984, 1985, 1989a, 1989b). Typically, in these studies, students are presented with passages or sentences containing errors, omissions, or inconsistencies within the text, and are asked to identify aspects of the text that make it difficult to understand. Students who detected more problems were considered better comprehension monitors than students who detected fewer problems. This method, although not typical of the kind of texts or reading situations students usually encounter, allows the researcher more direct behavioral evidence of students' monitoring than is provided by self-report measures.
Think-Aloud. Researchers have also examined monitoring with think-aloud or interview methodologies. For example, Pressley and Afflerbach (1995) summarize the results of a number of studies that have used think-aloud protocols to examine what students do as they read various types of texts. Consistently, these studies indicated the overall importance of monitoring in reading behavior; students who are better monitors of their reading show higher levels of reading comprehension and more learning. Further, these studies have identified a number of different aspects of monitoring including monitoring of the text characteristics, monitoring of self-understanding and problems in comprehension, and monitoring of cognitive processes used to read and understand text (Pressley & Afflerbach, 1995).

Given these different methods for measuring metacognitive judgments and monitoring, there are a number of issues to consider in terms of construct validity. We now turn to an analysis of the construct validity of these different measures of metacognitive judgment and monitoring.

Content Component. In the self-judgment methods of Nelson, Narens, Tobias, and others, individuals are asked to rate a set of items and then these same items are used in the performance or criterion task. In this sense, the internal logic for the study insures a perfect match, or overlap in terms of content representation, between the judgment task and the criterion task. On the other hand, the items used in these judgment tasks sample only a small range of possible content areas such as word definitions or arithmetic problems, leaving many other content areas not represented. Accordingly, if the judgment task samples the student’s awareness of vocabulary word definitions (see Tobias, this volume), this measure of monitoring does not necessarily represent the student’s monitoring of their mathematics knowledge.

In terms of meaningfulness, the reading and mathematics tasks used by researchers like Tobias, Baker, and Pressley are seemingly more relevant for academic learning than the paired-associate memory tasks used in the work by Nelson and Narens and their colleagues because of their similarity to classroom tasks. Nevertheless, some of the tasks used in the studies of reading have used texts with purposely misspelled words, nonsensical sentences, or other types of errors embedded in the text. Although these kind of tasks may be motivating and interesting for some students, like doing a puzzle or game, they are not representative of the usual texts students encounter in the classroom (e.g., textbooks) or outside the classroom (e.g., newspapers, magazines) that are designed to be error-free. It remains
an empirical question whether this difference in meaningfulness between authentic texts and “error-filled” texts influence students’ monitoring processes.

The research on metacognitive monitoring also illustrates how an assessment technique might adequately cover a broad spectrum of content within a particular domain but not across domains. Pressley and his colleagues, for instance, have examined monitoring while students read individual words and sentences, extended passages on the PSAT and SAT, or introductory psychology textbooks (Hunter-Blanks et al., 1988; Pressley et al., 1987a; Pressley et al., 1987b; Pressley et al., 1990). In a similar manner, Baker (1979, 1984, 1985, 1989a, 1989b) has examined students’ ability to monitor the presence of many different types of errors within different text formats. Hence, within a specific domain, such as reading comprehension, these researchers have used an array of content in their measures of monitoring. Similar results of students’ monitoring in other domains (e.g., science, social studies, reasoning, problem solving) would add to the evidence supporting inferences about the construct validity of our instruments. In addition, students’ prediction accuracy or error detection while completing a science experiment or while listening to a discussion in social studies would provide further insights into students’ overall monitoring behavior.

Substantive Component. Unlike the measures of metacognitive knowledge presented earlier, many of the measures of monitoring we have discussed in this section have not used items that could be readily subjected to factor analytic studies as a means of examining the internal relations among the item responses. However, in the comprehension monitoring research by Baker and Pressley, there is some evidence to suggest individual differences in detection of errors/text problems and monitoring of comprehension. Baker (1985,1989b) notes that there seem to be at least seven different types of standards that individuals can use to evaluate text, ranging from a lexical standard focused on individual word comprehension to more molar standards involving internal consistency and structural cohesiveness within the text. Individuals who use different standards will detect different errors or problems in the text. If the experimenter counts only detection of word errors, but a subject is using a more molar standard such as internal consistency or structural cohesiveness and does not detect the word errors, this subject may be considered a poor monitor, when in effect she is monitoring the text in a different manner.

As noted above in the metacognitive knowledge section, the issue of grain size and theoretical divisions of metacognition versus the
empirical evidence or resolution power of our instruments to adequately measure these divisions is important. On the one hand, much of the think-aloud literature reported by Pressley and Afflerbach (1995) suggests that monitoring and regulating processes often occur together and are difficult to separate empirically. On the other hand, there are good theoretical reasons for discussing monitoring and regulating as distinct processes (see Baker, 1989b; Zimmerman, 1989). This problem has implications for the development of self-report measures of monitoring. Developers of these types of measures may have to consciously choose whether to have measures that represent monitoring and regulation as relatively distinct aspects of metacognition and self-regulated learning thereby reflecting theory, or measures that blur the boundaries between these two components, reflecting much of the empirical evidence.

**Structural Component.** The relation between test scores and the construct of interest is of particular concern when considering metacognitive judgments and monitoring. Measures of EOL, JOL, and FOK rely on an analysis of the consistency between the subjects' responses to the judgment task (their perceptions) and their actual performance on the task. In the typical case, the pattern of responses can be organized in a two-by-two matrix representing the crossing of the judgments (yes/no about whether subjects know an item or not) with actual performance (yes/no regarding their recall or correctness of response). In this simple matrix, scores that are in the two cells of yes-yes and no-no are often called "hits" and reflect accuracy or calibration because of the match between judgment and performance. That is, the subjects judge they know it and they do (the yes-yes cell) or they judge they do not know it and they do not (the no-no cell). Subjects who have more scores in these two cells than the two off-diagonal cells (often called "misses" in judgment) are deemed to be better at monitoring or better calibrated given that there is substantial agreement between their judgments and actual performance. Subjects whose scores fall primarily in the yes-no cell (say they know it, but do not know it, an overestimation) or in the no-yes cell (say they do not know it, but then do recall it or know it, an underestimation) are assumed to be less effective at monitoring or less calibrated given the minimal agreement between their judgments and actual performance (cf. Tobias, this volume).

Although the methodological issues with this type of scoring system are complex and beyond the scope of this paper, we briefly mention one important consideration. Schraw (1995) calls attention to the distinction between measures of association and measures of
accuracy in developing scoring systems for analyzing the pattern of scores in the matrix of hits and misses. In his discussion, he points out that many studies have used gamma as a measure of association between judgments and performance scores. However, gamma reflects degree of association and not level of agreement. Using both mathematical and theoretical arguments, Schraw (1995) also shows that a simple matching coefficient does not capture all the information about accuracy either because it does not take into consideration miscalculations or mismatches (see Tobias, this volume; Tobias & Everson, 1995). For a measure of accuracy, he suggests the use of the Hamann coefficient, which includes information from both matches and mismatches by the student, thereby expanding the range of information that is used. He concludes that judgment studies should include measures of association like gamma as well as measures of accuracy such as the Hamann coefficient. Using an array of these types of measures of both association and accuracy will provide interval scales for data analysis and also will avoid the problem of using two simple measures of hits (or matches) and misses (or nonmatches) that are not independent from one another (i.e., if one has a high "hit" rate, then one's "miss" rate will be lower).

Another issue regarding the scoring of data from the matrix of hits and misses concerns the categorization of individuals into different groups, reflecting a more idiographic analysis. For example, scores from the judgment-performance relational data can be used to classify subjects into those who are calibrated (high agreement between judgments and performance), those who are overestimators (relatively high level of confidence in judgments and low level of performance), or those who are underestimators (low level of confidence in judgments and high level of performance). This type of scoring classifies individuals into three general groups in terms of their overall level of calibration. In the same way, Baker (1989b) suggests that there may be stable individual differences in reading comprehension monitoring resulting in two basic groups of skilled and unskilled readers. Again, this would reflect a more person-centered analysis focused on classifying students into two general groups of skilled and unskilled readers, or at least skilled and unskilled monitors of reading. Pressley and Afflerbach (1995) also suggest that some type of categorical system that distinguishes between good and poor readers may capture much of the important variance. This type of categorical analysis conceptualizes metacognitive monitoring in terms of different "types" of people who are either good or poor monitors, rather than the idea that individuals can and
do vary along a continuum in terms of their monitoring ability. Accordingly, a model that proposes that the construct of monitoring should be represented along a continuum should generate and use the various continuous measures of monitoring discussed above. In contrast, a more person-centered model of monitoring that stresses a disjunction between good and poor monitors should generate and utilize dichotomous scoring methods.

**External Component.** Questions of how the various measures of metacognitive judgment and monitoring are related to: (a) each other, (b) measures of metacognitive knowledge and regulation, and (c) other constructs such as prior knowledge and general intelligence are addressed under the external component of construct validity. In the metamemory research on EOL, JOL, and FOK measures, Leonesio and Nelson (1990) have shown that these three types of judgment measures are only weakly related to one another. Correlations ranging from .12 to .17 among the three measures suggests that EOL, JOL, and FOK judgments are tapping different aspects of monitoring. Pressley and his colleagues (e.g., Hunter-Blanks, Ghatalla, Pressley, & Levin, 1988; Pressley, Snyder, Levin, Murray, & Ghatalla, 1987b) examined the relations among various measures of monitoring by using judgments of learning at different times during the reading-testing process, reflecting EOL, JOL, FOK, and confidence measures of monitoring. Results indicated that JOLs and confidence ratings were more closely tied to performance in comparisons to EOLs or FOKs. In most of the studies by Pressley and his colleagues, students were assigned to one of three conditions defined by when they were asked to make their judgments (before reading, after reading, or after testing). This type of between-subject design does not allow for comparisons within individuals across measures as in the Leonesio and Nelson (1990) study. Accordingly, although there is experimental evidence that different types of judgments (i.e., EOLs, JOLs, and FOKs) can have different relations to performance, thereby suggesting different functions for these components (Nelson, 1996), there is still a need for within-subject designs that allow for intraindividual comparisons of the relations among EOLs, JOLs, and FOKs.

In terms of how monitoring is related to metacognitive knowledge and control or regulation, the findings are mixed. Pressley and Afflerbach (1995) have shown that monitoring and regulating are often reported together in think-aloud protocols. Paris and Oka (1986) have shown that metacognitive knowledge is weakly related to performance on error detection tasks with correlations ranging from .15 to .30. Baker (1989b) notes that predictions of learning (EOLs),
judgments of comprehension or learning (JOLs), and confidence in learning (postdictions of learning) are often not clearly related to performance. This type of mixed evidence signals the need for research to clarify the conceptual relations among the three general components of metacognitive knowledge, monitoring, and control and regulation as well as their relations with actual performance.

Finally, measures of monitoring should assess monitoring and not other constructs such as verbal ability, prior knowledge, or general intelligence (Pressley & Ghatala, 1990). In the error-detection method described above, students are told to look for errors. This may invoke a level of monitoring in which the students do not typically engage when reading. Accordingly, performance on the error detection task may not represent spontaneous monitoring, the actual construct of interest. Moreover, students’ monitoring per se is not measured, rather monitoring is operationalized as the reporting of problems in the text. In addition, students may notice but hesitate to report problems with the text because they fear being wrong or because of epistemic beliefs about text that constrain their reporting of errors. For example, as Baker (1989b) has pointed out, if students endorse the cooperative text principle of Grice, they generally believe that texts are correct and should be error-free. When operating with this belief, students will be unlikely to report all errors in the text. To the extent that factors other than metacognitive monitoring influence students’ reporting of problems in the text, results of error-detection studies may be challenged in terms of the evidence they provide of construct validity.

In the think-aloud studies, spontaneous monitoring is evoked and can be assessed quantitatively and qualitatively (Pressley & Afflerbach, 1995). However, think-aloud protocols require students to perform the actual task and simultaneously verbalize their thoughts. The cognitive demands of this dual task may vary with the expertise or knowledge of the student, the extent to which students have automatized some of the cognitive activities, the age of the student, and/or their verbal ability. Consequently, verbal reports of monitoring may be confounded with these other constructs and may not provide the best evidence for construct validity.

Unlike the work on metacognitive knowledge, there is not as much validity research on how students’ monitoring is related to performance on standardized achievement tests. Given that monitoring is an ongoing process for a specific text or task, the relation to standardized general achievement tests may be variable. For example, Pressley et al. (1990) found that actual comprehension perf-
mance was moderately correlated with scores on a subset of SAT verbal items (rs ranged from .42 to .59), but that these same scores were not significantly related to monitoring of comprehension as measured by prediction scores (rs ranged from -.24 to .39, but were not significant due to power of test, small sample sizes). Pressley et al. (1987b) also reported no significant relation between students' general abilities (as measured with SAT and GRE items) and their monitoring as measured by estimates of test readiness (a JOL judgment). Paris and Oka (1986), however, did report that measures of error detection were positively related to performance on a standardized reading test, even after general intelligence was partialled out, although the magnitude of the relation was small (rs ranged from .09 to .23). Tobias and Everson (1995) reported that their measures of metacognitive judgments of mathematics knowledge were highly correlated with scores on the mathematics section of the Metropolitan Achievement Test (r = .76 for correct estimate scores, and r = -.72 for incorrect estimates scores). Despite these encouraging results, more research on how various measures of monitoring are related to standardized measures of achievement and ability will improve our understanding of the relations among the different aspects of monitoring and other constructs such as general ability, achievement, and learning.

Generality of Construct Meaning. Research on measures of monitoring have been carried out with groups of students varying in age and gender but not varying in ethnic/racial/cultural background. For example, Pressley et al. (1987a) included students from the first and fifth grade, Pressley and Ghatala (1989) included first, second, fourth, fifth, seventh and eighth graders, whereas Pressley et al. (1987b) and Hunter-Blanks et al. (1988) used undergraduate students. Baker has focused on adults and much of the metamemory research of Nelson and Narens and their colleagues has been carried out with college students. Although studies by Baker (see summary in Baker, 1989b) have tended to rely on the adult reader, students of different developmental levels have also been included. In addition, Baker has examined how error detection might differ among good and poor readers. In some of these studies gender differences were explored and found (e.g., Pressley et al., 1987a; Pressley & Ghatala, 1989). Hence, results from this work extend over a broad age range and across gender, suggesting the generalizability of the measures across diverse populations. Nevertheless, like much of the research in psychology and education (see Betancourt & Lopez, 1993; Graham, 1992), there is a large void in terms of our understanding of these constructs and measures in diverse cultural, racial, and ethnic populations.
Besides this issue of sample generality, there remains the perpetual and crucial issue of domain generality. Does a general monitoring skill exist or is monitoring dependent on domain expertise or other personal and contextual factors (Schraw, Dunkle, Bendixen, & Roedel, 1995)? Schraw et al. (1995) have shown that there are domain-general and domain-specific aspects of metacognitive monitoring. They found that confidence judgments were correlated across different domains of knowledge and a factor analysis of several different measures of monitoring (confidence, discrimination, bias) did generate one general monitoring factor. However, they also found that a measure of monitoring accuracy (the discrimination score, which takes into account correct and incorrect predictions) showed a domain-specific pattern of results. They suggest that there may be a developmental progression from a domain-specific expertise to a general monitoring skill. Accordingly, measures of monitoring have to be sensitive to developmental and domain-specific factors that might bear on the construct validity of domain-general measures of monitoring.

Relevance and Utility Concerns for Test Use. As noted above in the metacognitive knowledge section, measures vary in their relevance and utility for researchers versus practitioners. Think-aloud protocols offer a window into the kinds of monitoring processes that individuals use as they perform cognitive tasks and are therefore probably best suited for researchers who are attempting to provide a detailed description of the various monitoring processes (Pressley & Afflerbach, 1995). However, protocol analysis is time and labor-intensive, requires specific training, and cannot be used easily or efficiently with large groups of students thereby limiting its use in classroom settings. In contrast, methods like the error-detection tasks can be used by both researchers and teachers. These tasks do not rely on verbal reports and can be tied closely to the types of classroom tasks in which many students engage on a regular basis such as reading different texts. As such, they provide opportunities for teachers to examine monitoring in their students quickly and to guide lessons on reading comprehension skills.

Formal measures of EOL, JOL, FOK, and confidence are probably best suited for use by researchers. It seems unlikely that teachers would have students go through and rate all the items on a test or questions on a worksheet in terms of their difficulty before, during, and/or after performing the task. However, teachers can use informal methods of assessment by asking students to think about their prior knowledge before reading a text, or self-question themselves
about their understanding (a good strategy for monitoring) during reading, or self-test themselves after reading a text. These informal assessment procedures may be useful to teachers, at a very global level, to determine students' ability to monitor their comprehension.

**Summary.** There are a number of instruments that can be used to assess metacognitive judgment and monitoring skills. Although each of the measures discussed in this section is backed by considerable empirical data, there remain a number of unresolved issues. First, at a conceptual level, researchers need to be careful in terms of their labels for different aspects of metacognitive judgments in terms of EOLs, JOLs, FOKs, and confidence ratings. The proliferation of different labels for the same basic constructs makes it difficult to summarize and compare results from different studies. If any science is to make significant advances, there is a need for clearly defined and agreed-upon labels for the constructs under study. The area of metacognitive monitoring has generated a large number of different terms that do not help facilitate communication. We propose that the framework of EOLs, JOLs, FOKs, and confidence judgments is a reasonable start in the direction of fostering clarity and consensus. Second, because current measures are primarily restricted to reading and mathematics, questions regarding the operation of monitoring in other domains remain unanswered. Third, there remains an issue about the conceptual and empirical separation of monitoring from regulating. As we saw in the previous section on metacognitive knowledge, our theoretical models of metacognitive monitoring propose more distinctions and subcomponents than what is often found in our empirical data. In particular, most models propose a separation of monitoring and regulation, but these two components are often fused in learning. It may be difficult to develop assessment instruments that can reliably and validly tease these subcomponents apart, so our conceptual models may have to be adjusted unless we can develop instruments with higher resolution power. Fourth, in terms of scoring measures of metacognitive judgments (structural component), more careful measures of both association and accuracy need to be used as Schraw (1995) has proposed. Fifth, although the evidence for the external component of construct validity is fairly good, there is still a need for more research with diverse samples to improve the generality of construct meaning. Finally, efforts are needed to bridge the gap between experimental methods of assessing metacognitive monitoring (e.g., EOLs, JOLs, etc. from metamemory paradigm; Nelson & Narens, 1990) as well as the error detection tasks and think-aloud protocols and those (e.g., ratings of knowledge; see
Measures of Self-Regulation and Control of Cognition

Although there are data suggesting that monitoring and regulation are often fused in actual performance (Pressley & Afflerbach, 1995), measures have been developed that focus more on regulation and control of cognition than on monitoring. Three general methods have been used to assess regulation: think-aloud protocols, self-report questionnaires, and interviews. We have already described think-aloud protocols in the previous section. Here we concentrate on self-report questionnaires and interviews. A number of different questionnaires have been used to assess various aspects of regulation including the Learning and Study Strategies Inventory or LASSI (Weinstein et al., 1987; Weinstein et al., 1988), the Motivated Strategies for Learning Questionnaire or MSLQ (Pintrich & De Groot, 1990; Pintrich, Smith et al., 1993) and other more focused instruments such as Kuhl’s action-control scale (Kuhl, 1985) and other study skills instruments such as Brown and Holtzman’s (1967) Survey of Study Habits and Attitudes. Given our research with the MSLQ, we concentrate on that instrument as representative of a questionnaire to measure self-regulated learning. For comparative purposes, we consider some aspects of the LASSI.

Self-regulated learning has also been measured in various interview studies, but the Self-Regulated Learning Interview Schedule (SRLIS) developed by Zimmerman and Martinez-Pons (1986, 1988) is the most formalized interview measure available, so we concentrate on this exemplar in this section. Together, these three instruments—the MSLQ, the LASSI, and the SRLIS—can be used to illustrate some of the important construct validity issues concerning assessments of students’ regulation of their cognition.

The MSLQ and LASSI are self-report instruments that ask students to respond to Likert-type items concerning their level of cognitive strategy use and their regulation of cognition. The key difference between the two instruments is the theoretical assumption about the nature of self-regulation underlying their development. The LASSI was developed from a domain-general perspective. Students are asked about what they do in general in terms of their learning. The MSLQ reflects a more domain-specific view, at least in terms of domain specificity being operationalized at the course level. Students are asked to respond to the items in terms of what they do in a specific course or class. The MSLQ is not task specific (e.g., exam, reading
textbook, writing a paper) or knowledge-base specific (biology, mathematics, history, etc.), which might be important from some perspectives (see Schraw et al., 1995).

In terms of cognitive strategy use, individual scales on the MSLQ are designed to assess rehearsal, elaboration, organization, and critical thinking, whereas metacognitive monitoring and self-regulation are assessed using one 12-item scale (Pintrich, Smith et al., 1993). In addition, resource management strategies are assessed in four different scales, including time and study management, effort regulation, peer learning, and help seeking (Pintrich, Smith et al., 1993). A typical question from the regulation scale of the MSLQ states, "When I become confused about something I'm reading for this class, I go back and try to figure it out."

The SRLIS, using an individual-interview format, asks respondents about specific tasks with follow-up probes questioning how they would behave in six different academic contexts (Zimmerman & Martinez-Pons, 1986). These contexts are a classroom discussion, short writing assignment, mathematics assignment, end-of-term test, homework assignment, and studying at home. Students are presented with a one- or two-sentence description of the context and then asked about their methods for managing the situation (Zimmerman & Martinez-Pons, 1986). For example, for the test-taking context students are told, "Most teachers give a test at the end of the marking period, and these tests greatly determine the final grade." Then, they are asked "Do you have a particular method for preparing for a test in classes like English or history?" Whereas ratings of the items on the MSLQ are averaged into scales, SRLIS responses are categorized into 1 of 14 different categories representing knowledge (e.g., organizing), monitoring behavior (e.g., keeping records, self-evaluation), strategy use (e.g., rehearsing and memorizing), and regulation (e.g., goal setting and planning).

Content Component. As reflected in the many scales of the MSLQ and the 14 different categories of strategies from the SRLIS, items on these measures attempt to cover important content in self-regulation and control of cognition. There is also evidence that these two measures strive to represent content from the diverse domain of regulation strategies by sampling strategies related to many different academic activities. The MSLQ queries students about one particular class and focuses on reading and study activities, although a few items refer to other academic situations (e.g., note taking, listening to lectures). In other research, however, items on the MSLQ have been modified to cover specifically a broader range of academic contexts.
by preceding items with cues to different situations such as "When I study for a test . . .," "When I do homework . . .," or "When the teacher is talking . . ." (Pintrich & De Groot, 1990). More recently, items from this instrument have been used specifically to assess regulation within different subject areas in order to evaluate between-domain differences (Wolters & Pintrich, 1998).

The SRLIS also asks students to report their strategy use across a variety of academic tasks (e.g., classroom discussion, test taking, and homework) and different academic subject areas (e.g., history, mathematics). Further, this open-ended interview allows students freedom to respond with the particular strategies they use in these different contexts. Overall, both of these measures provide a breadth of coverage in terms of strategies for different tasks and subject areas as well as the type of strategies assessed. This coverage seems to provide reasonable content representativeness of the many different control and regulation strategies available relative to other assessments that focus on one type of task, one academic domain, or a small number of strategies.

**Substantive Component.** The MSLQ provides a reasonable match between the theoretical model and the empirical results of confirmatory factor analyses with data from college students (see Pintrich, Smith et al., 1993). For example, in our structural equation models we have a chi square ratio of 2.26 (values under 5.00 are considered optimal), a GFI of .78 (GFIs of .90 or above are considered optimal), and a CN of 180 (CNs of 200 and above are considered optimal). Although some of the fit statistics for our structural equation models could be improved by having a less theoretically based factor structure, we have opted to maintain the theoretical structure as long as the data provide a reasonable fit to the model. Of course, this problem of lack of a stronger fit between the theoretical model and the actual empirical data parallels the problems mentioned previously in our discussion of both metacognitive knowledge and metacognitive monitoring. In general, the problem remains that our conceptual models propose more components and complexity than are supported by the empirical data.

Using data from younger students, such as junior high school students, we have not been able to reproduce as detailed a factor structure as in the college data (Pintrich & De Groot, 1990). For example, rather than three scales that reflect different types of cognitive strategy use, the junior high data only formed one scale reflecting students' combined use of rehearsal, elaboration, and organizational strategies. In the same fashion, the two scales of metacognitive and
effort management, distinct in the college student data, combined into one scale with the junior high students. These results could reflect the general developmental orthogenetic principle of Werner (1948), which suggests that, with development, systems change from being relatively undifferentiated to having very differentiated components organized into a hierarchy. On the other hand, the results could just reflect a problem in generality of construct meaning with younger students or a problem with construct irrelevant variance arising from the use of self-reports with young students.

In both the college and junior high data, there was no support for separate metacognitive scales of planning, monitoring, and regulating. Hence, although the underlying theory suggests that these aspects of metacognition and self-regulation should be distinguishable, the data do not support this assumption. Results such as these challenge the substantial component of construct validity and highlight the grain size and instrument resolution problem mentioned previously.

The SRLIS produces interview data coded according to 14 different categories of strategies that are based upon a specific theory of self-regulation (see Zimmerman, 1989, 1994). On one hand, given that these categories determine the type of information extracted from students’ interviews, this instrument may have a higher degree of substantive validity than interviews that code responses using post hoc categories. On the other hand, Zimmerman and Martinez-Pons (1988) have shown that a principal component analysis groups 12 out of 14 categories into one large factor that they call Student Self-Regulated Learning. Again, paralleling the data from the IRA and MAI on metacognitive knowledge, the think-aloud protocol data from Pressley and Afflerbach (1995) on metacognitive monitoring, and the MSLQ data on regulation, these findings from the SRLIS suggest that students who engage in one component of self-regulated learning also engage in other components. Accordingly, efforts to separate the different components into theoretically smaller subcomponents may not be justified by the empirical data.

Taken together, the data from the MSLQ and SRLIS, as well as the monitoring data from the think-aloud protocols, suggest that although we can distinguish monitoring and regulation theoretically, the empirical data are more ambiguous. It appears that some students tend to engage in a variety of these strategies and other students are less likely to report using them. There is clearly a need for more specification of the theoretical model or nomological network of constructs that involve both monitoring and regulation, followed by
careful research on how these improved models can help us develop substantively valid and high resolution measures.

**Structural Component.** One difficulty with data from the SRLIS is that it is not easy to quantify the scores in a manner that will yield interval data. Zimmerman and Martinez-Pons (1986), for example, proposed and tested three different methods for scoring results from the SRLIS. Based on its ability to distinguish between students of different achievement groups (using the Metropolitan Achievement Tests), Zimmerman and Martinez-Pons (1986) chose a scoring method dependent on students' mention of a particular strategy and their report of how often they used that strategy. This overall measure seemed to provide a better index than did counts of strategy use or strategy frequency, although these other two measures also discriminated between the two achievement groups.

The MSLQ is scored by taking the mean of the students' ratings for the items that comprise a scale. However, it should be noted that the MSLQ does not provide any normative data for comparison as does the LASSI. Users of the LASSI have available the norms for a large sample of students and comparisons can be made between an individual's score on a scale and the scale score based on the normative sample. In contrast, the MSLQ is based on the assumption that students' use of strategies and self-regulation may vary by type of course and specific classes and so norms are not provided. Although this may be more in line with current views of self-regulated learning, lack of normative data restricts some of the practical uses of the MSLQ.

Given the differences in scoring and conceptual models, there may be some evidence to support the use of more idiographic or person-centered categorical systems of scoring that simply classify students into good or poor self-regulators or strategy users (see Pressley, Harris, & Marks, 1992). The use of norms, as in the LASSI, suggests that students can be compared and then classified into different categories of more or less self-regulating learners. The interview data from the SRLIS could also be used in such a manner. As was discussed in the section on monitoring, this distinction between continuous versus categorical scoring systems is an important one for future research to address.

**External Component.** One issue with the MSLQ and the SRLIS is that, like the assessments of monitoring, it is not clear if the measures primarily assess the construct of interest. Both of these instruments ask students to report, retrospectively, how they behave in general types of situations. These measures do not question students about a previously completed specific task. Because of this format, students
are more apt to access long-term memory and make generalizations about what they believe they do in a particular situation. Consequently, self-reports have been criticized for their potential to be biased or inaccurate (Ericsson & Simon, 1993; Garner, 1988; Pressley & Afflerbach, 1995). For instance, on the MSLQ students may endorse the statement “I try to change the way I study in order to fit the course requirements and instructor’s teaching style,” not because they really change their study behavior, but because they know that this would be a good strategy. In this case, student responses may tap into metacognitive knowledge as well as regulation of cognition. One way to remedy this problem is to adapt items so that students are referring to specific recent incidents or recently completed tasks when they respond. Nevertheless, these self-report measures are still subject to problems of students having conscious access to their strategy use, being able to verbalize their strategy use, as well as being unbiased and accurate in their reporting (i.e., reporting more strategy use than actually engaged in for social desirability reasons).

Another way to address this problem would be to assess control and regulation using a more “on-line” methodology such as stimulated recall or think-aloud. Howard-Rose and Winne (1993), for example, devised a task in which students reported what they were thinking and doing while they were still in the process of reading a passage, perhaps giving a more direct measure of regulation and monitoring. The considerable time and effort involved in instructing students in how to perform this task combined with the actual administration time, reduce the utility of this method. Furthermore, the ability to generalize this task to other tasks and other populations may be limited to older students who are able to manage the cognitive load produced by simultaneous task completion and the think-aloud task.

Although these problems of construct validity are always present with self-report instruments, there is evidence bearing on the external component of construct validity that supports the use of self-report instruments. Students’ scores on the regulation portions of the MSLQ and the SRLIS have been linked in predictable ways to a number of other indicators of learning, performance, and motivation. Strategy use and regulation, as measured by the MSLQ, were related to seventh-grade students’ first and second semester grades and their achievement on different types of classroom tasks where regulation may influence performance (Pintrich & Garcia, 1991; Pintrich & De Groot, 1990). For example, Pintrich and De Groot (1990) found that the MSLQ regulation scale was related to students’ performance during seatwork, on tests or quizzes, and on report writing. Although correlations were not high (rs
range from .20 to .32), they do indicate some relation between academic performance and the regulation scales on the MSLQ. Scores on the strategy use and regulation scales of the MSLQ have been related, in theoretically predictable ways, to components of students' motivation including self-efficacy, task value, intrinsic motivation and test-anxiety (Pintrich, 1989; Pintrich & De Groot, 1990; Pintrich & Garcia, 1991; Pintrich, Roeser, & De Groot, 1994; Pintrich, Smith et al., 1993).

Whereas the SRLIS has not been linked to performance on specific classroom tasks like the MSLQ, Zimmerman and Martinez-Pons (1986, 1988) did find that students' responses to the SRLIS were related to teachers' ratings of students' efforts at regulation and to achievement on the Metropolitan Achievement Test. In short, positive efforts have been made to examine the expected relations between strategy use and regulation as measured by the MSLQ and the SRLIS and other constructs such as achievement, teachers' grades and ratings, and motivational constructs. Although the magnitude of the relations is modest, the evidence indicates that scores on these instruments, at a minimum, can distinguish between high and low achievers in classroom settings as predicted by our conceptual and theoretical models of self-regulated learning.

Generality of Construct Meaning. The generalizability of an assessment across different domains is an important aspect of construct validity. The MSLQ and the SRLIS have items that refer to distinct academic tasks and subject areas. Hence, results about students' ability to regulate their cognition may not be limited to a single domain as the case may be with instruments that reference only one domain (i.e., reading comprehension).

As we have noted, the diversity and size of the samples used in the studies of these measures are important to consider when assessing generality of construct meaning. Data for the MSLQ were initially collected with a fairly large number of college students (different samples of 326, 687, 758 and 380, for a total of over 2,000) from different types of institutions (research universities, comprehensive universities, small liberal arts colleges, and community colleges) spanning many different subject areas (see Pintrich, Smith et al., 1993). It has also been used extensively with middle school students (Pintrich & De Groot, 1990; Pintrich et al., 1994), but has not, to our knowledge, been used specifically to examine students below the seventh grade or special populations such as students with learning disabilities or gifted students. Self-report questionnaires may be difficult for younger children or those of lower achievement levels who may not be able to read the items on the questionnaire. Interviews or reading the
questionnaire items to students can help in this regard, but interviews may still be better with younger children. In terms of sampling issues, Zimmerman and Martínez-Pons (1988) used the SRLIS to examine metacognition and self-regulation in a relatively small sample (N = 80) of high and low track high school students. Clearly, as we have already noted for the measures of both metacognitive knowledge and monitoring, there is a need for research on these control and regulation instruments with more ethnically and racially diverse populations as well as students across a range of grade (age) and achievement levels.

Relevance and Utility Concerns of Test Use. One reason for the difference in the samples sizes of studies using the MSLQ and those using the SRLIS is likely the relative ease of administration of the MSLQ. Self-report measures, as exemplified by the MSLQ, can have relatively high degrees of utility value for research studies or more practical uses because they can be completed quickly and easily by large numbers of students. In addition, they can be used by teachers or researchers in classroom settings without much disruption of routines. One or two individuals can administer the questionnaire to large numbers of students over a relatively short time frame and the data collected are fairly easily transferred to analyzable form.

In comparison, even short interviews such as the SRLIS must be individually administered and therefore take substantially longer to complete. Further, the resulting data require a labor intensive effort to change into a usable format. Thus, one advantage to questionnaires such as the MSLQ or LASSI is that the researcher is able to collect a great deal of information quickly and easily. At the same time, open-ended interviews like the SRLIS have an advantage in that they allow students more freedom to respond because they do not limit responses to particular strategies. This aspect of these interviews may increase the relevance that scores have for more diagnostic purposes. The interview data can provide a good window into the students’ general schema for learning, a more “Gestalt” like view of their approach to learning and self-regulation that can get lost in the division of self-regulated learning into the multiple scales of the MSLQ.

Summary. Both self-report questionnaires and interview methods can provide reasonable measures of control of learning and self-regulated learning. First, these measures seem to provide good content representativeness of a number of different types of general strategies for control and regulation of learning, although they do not include domain-specific control and regulation strategies (e.g., a control strategy for math problem solving; a control strategy for writing an essay). Second, as noted in our discussion of measures of
metacognitive knowledge and monitoring, there are still major questions about the fit between the complex theoretical models (in terms of number of subcomponents) and the somewhat simple models supported by the empirical data. Third, the scoring systems are reasonable and easy to use, although there still remain questions concerning the use of continuous measures of self-regulation versus categorical scoring systems based on a simple dichotomy of good and poor self-regulators or strategy users (see Pressley, Harris, & Marks, 1992). Fourth, the major issue in terms of construct validity of self-report questionnaires or interviews concerns their susceptibility to problems of construct-irrelevant variance stemming from differences in individuals' ability to consciously access their strategy use and control efforts, verbalize their strategy use, read the questionnaire items, or their susceptibility to social desirability or other forms of bias. Much work needs to be done to resolve these problems with self-report questionnaires and interviews. Fifth, as with all the measures there is a great need for the use of more diverse samples. Finally, questionnaires can be used easily and quickly with large groups of students in classroom situations and can be a very practical alternative to more experimental methods. Interviews can avoid some of the problems of questionnaires in terms of construct irrelevant variance by the judicious use of probes and focusing the student on specific tasks, but they are more time-consuming and costly to use.

**GENERAL CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH AND DEVELOPMENT ACTIVITIES**

Given our discussion of various measures of metacognition and self-regulated learning, there are a number of conclusions that can be drawn from our review of the evidence for the construct validity of these measures. These conclusions suggest several fruitful directions for research and development activities.

1. There is no one “perfect” measure of metacognition.

As we have seen throughout this chapter, there are a number of different measures and methods that have been used. Oftentimes the strength of one method is the weakness of another. Certain methods, such as think-aloud protocols, although potentially supplying “real-time” measures of metacognition and self-regulated learning in given contexts are difficult to use on a large-scale basis. In contrast, self-report questionnaires are high in applied utility, but are open to criticism regarding the potential for construct-irrelevant variance to be generated by the self-reports. Individual researchers and practitioners must deter-
mine what their purposes and needs are and then make informed choices about what measures to use given their own goals and the context in which they are working.

2. Different instruments measure different components of metacognition.

Measures of metacognitive knowledge do not necessarily tap into aspects of monitoring or regulation. In fact, different measures of metacognitive judgment or monitoring do not even assess the same components of metacognitive monitoring and judgment. Consequently individuals must be clear on which aspect they are interested in and choose instruments that match their interest because the measures cannot be used interchangeably. To facilitate this type of informed decision making, researchers and instrument developers need to be clear about which component of metacognition their instrument assesses and label their instrument accordingly. We have proposed a general three-component model of metacognition and self-regulated learning that includes a number of important subcomponents. We think this model is grounded in current theory and research and should be helpful in clarifying which components of metacognition the various instruments are tapping. Instrument developers who use this three-component model of metacognition and self-regulated learning and label their instruments in line with it will ensure some consistency in assessment use and facilitate cross-study comparisons of empirical findings.

3. Further specification of the theoretical relations among the different components of metacognition and self-regulated learning would be helpful for instrument development.

As has been noted throughout this chapter, there is a disjunction between our theoretical models and the empirical data, particularly with respect to monitoring and regulation. Most models separate these two components and the separation makes sense intuitively and conceptually, but the empirical data argue against a separation. This is the problem of "grain size" and instrument resolution (see Howard-Rose & Winne, 1993). For some purposes, a general distinction between high and low levels of self-regulation (Zimmerman & Martinez-Pons, 1986, 1988), or good and poor strategy users (Pressley et al, 1992), or more or less cognitive engagement (Pintrich & Schrauben, 1992) can be fruitful. In other contexts and given other research goals, there is a need for more fine-grained analysis of the component and subcomponent processes. Theoretical and empirical work on these issues will clarify our models and help us develop more conceptually sound and useful instruments.
4. Construct validity studies are needed to test our theoretical models and the validity of our assessment instruments.

Given that metacognition and self-regulated learning include a number of different components and that there are a number of different methods that can be used to assess these components, there is a need for careful and well-designed construct validity studies. For example, multitrait, multimethod (MTMM) studies can be used to clarify our theoretical models as well as provide us with useful information about our instruments. The recent MTMM study by Howard-Rose and Winne (1993) on self-regulated learning demonstrates how a MTMM study can help to clarify our conceptual models as well as suggest how best to measure different components of metacognition and self-regulated learning. There have been very few carefully done studies like this in the area of metacognition and self-regulation and more MTMM studies would certainly benefit the field. At the same time, we can go beyond the correlational analyses of MTMM studies and examine the different factors that contribute to the variance of our measurement instruments through the use of generalizability studies. For example, generalizability studies (see Baxter & Shavelson, 1994; Gao et al., 1996) can provide data on the comparability of different tasks and methods for assessing the different components of metacognition and self-regulation.

5. One of the most problematic issues from both theoretical and measurement perspectives is the domain specificity vs. generality issue.

Metacognition and self-regulated learning are generally measured with respect to one domain such as reading comprehension, but they are often considered domain-general constructs that transfer or generalize across domains. For example, it is often assumed that students who are high in metacognitive monitoring or general self-regulation for one task will also be able to transfer these skills to another task or domain. In terms of content representativeness, many of our measures have focused on the content areas of memory or reading comprehension. There is a need for more research and development in other academic areas such as mathematics, science, and social studies. In addition, in our measures of metacognition and self-regulated learning, we need to address how individual scores on our instrument generalize or transfer across domains (see Schraw et al., 1995). Our theoretical models have not always been clear concerning how transfer is assumed to occur across situations, tasks, or domains, so it is not surprising that our measurement efforts have been less than successful in coping with this issue. The issue of domain specificity and transfer may be the largest and most intractable problem confronting our theoretical and assessment efforts.
6. The use of performance assessments may help us measure both knowledge and metacognition within and across domains.

Recent developments in the use of performance assessments may help us resolve the tension between knowledge-based or domain-specific models of learning and cognition that focus on students’ prior knowledge and more domain-general models that stress the role of metacognition and self-regulated learning. For example, Baxter, Elder, and Glaser (1994, 1996), have examined performance assessments designed to provide measures of students’ knowledge in science domains (life science, physical science). These performance assessments use tasks that are meaningful and relevant to students and are typically used in classroom settings to monitor instruction. Interviews of students while carrying out the assessment provided evidence of general monitoring and regulation strategies. Moreover, students who performed well on the science assessment displayed more frequent and flexible monitoring strategies than did students who performed less well. Research that attempts to examine the use of metacognitive strategies in everyday classroom contexts and how these strategies relate to performance within and across tasks sheds important light on our understanding of metacognition and suggests how instructional changes might be implemented to enhance learning in the classroom.

7. There is a need for longitudinal research across ages.

Cross-sectional studies of different aged students show that metacognition develops with age and experience, but we have relatively few studies that show metacognitive development within individuals. We need studies to test the theory that children first develop domain-specific metacognitive knowledge or strategies, followed by a more generalized transfer of these strategies to a number of domains (see Schraw et al., 1995). Moreover, these kinds of studies can guide instrument development and perhaps lead to different types of measures being used at different ages.

8. There is a need for research with diverse populations.

Obviously, as we have pointed out throughout this chapter, there is a need for more research and test development activities that include diverse ethnic and racial groups. Although most models of metacognition should be applicable to all groups of individuals, there is some evidence that different groups of students may make judgments about themselves in somewhat different ways. Graham (1994) points out that many African-American students’ perceptions of their learning and confidence in their ability are not highly correlated with their actual achievement scores or performance. As we have discussed in this paper, this is a problem in the calibration of monitoring judgments and actual perfor-
mance. Graham (1994) notes that it is not clear theoretically why this may be the case, but there have been suggestions that this is an adaptive coping strategy given that many of these students have generally low academic performance. If this explanation is correct, then it suggests that for these students, they may be making poor judgments of their learning and understanding in order to maintain their motivation and self-beliefs. However, this poor calibration can have detrimental effects on the use of regulating strategies. If the students believe that they are learning, when they are not, then they will be unlikely to change or effectively regulate their cognition and learning.

This type of dynamic is not addressed in most of the research on metacognitive monitoring and self-regulated learning and we need to test our models with diverse groups of students to determine if there are different processes involved for some of these groups. To the extent that there are different processes for these groups, and our normative models of metacognition and self-regulated learning do not include these processes, then this can result in instruments that suffer from construct underrepresentation of these different processes for diverse groups. Alternatively, our current models and instruments may suffer from problems of construct-irrelevant variance if these different groups respond to our instruments in a different manner than what is predicted by our normative models.

In summary, our models and instruments are developed to the point that they are useful for field work and the improvement of practice. At the same time, there is much theoretical and empirical work to be done in the area of metacognition and self-regulation to clarify our models and substantiate the "adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1989, p. 13). We hope that the discussion in this chapter will stimulate researchers in the field to continue to question their instruments in an effort to improve our assessment methods and build our understanding of the nature of metacognition and self-regulation.

REFERENCES


2. ASSESSMENT OF METACOGNITION


Assessing Metacognition in Children and Adults

Linda Baker

University of Maryland, Baltimore County

Lorraine C. Cerro

University of Maryland, Baltimore County

It has been about 25 years now since researchers first became interested in the study of metacognition, with the onset of interest marked by the publication of the 1975 metamemory interview study of Kreutzer, Leonard, and Flavell and the seminal theoretical work of John Flavell (1976) and Ann Brown (1978). The early work by developmental psychologists on age-related differences in children’s metacognition captured the attention of researchers concerned with individual differences in academic achievement in children as well as adults. Within academic domains, most of the research has been focused on reading and studying (Baker & Brown, 1984; Forrest-Pressley & Waller, 1984; Garner, 1987; Paris, Wasik, & Turner, 1991), but mathematics (Van Haneghan & Baker, 1989), writing (Scardamalia & Bereiter, 1985), and science (Baker, 1991) have also received attention. The consistent finding has been that students who are more successful in a domain exhibit higher levels of metacognitive knowledge about the domain and are more skilled at regulating their cognitive processes.

Clearly, the construct of metacognition has had wide appeal and wide applicability, stimulating a great deal of research across a broad
spectrum of psychological problems and issues, as well as a growing amount of intervention work in classrooms. In a 1994 review paper on social influences on metacognitive development, Baker wrote, "The popular appeal of metacognition has led to the widespread adoption and somewhat uncritical acceptance of the construct among educators. This situation is obviously problematic from a scientific standpoint and makes clear the need for further basic research on how metacognition develops, the role of metacognition in cognitive development, and how metacognition may best be fostered" (pp. 202-203). The concern about uncritical acceptance is no less apt with regard to measurement; let us therefore amend the final sentence to end with and measured.

In this chapter, we address the issue of metacognitive assessment first by examining methods of measuring metacognition used in empirical research, including questionnaires, interviews, think-aloud procedures, error-detection procedures, and various on-line measures. We then examine some of the instruments that have been subjected to tests of reliability and validity by independent investigators; their numbers are few. Next we consider recommendations for assessing metacognition that are published in books and journals for teachers and school psychologists; their numbers are many. Throughout, primary emphasis is on metacognition as it relates to reading and studying, but some reference is made to assessment of metacognition in other domains as well (e.g., metamemory, problem solving).

The literature focusing specifically on metacognitive assessment is sparse, but many researchers have discussed issues related to assessment in their own empirical investigations as they seek to justify the measures they have chosen. In addition, much relevant writing appears in papers on the assessment of reading or academic achievement in general rather than the assessment of metacognition per se. We will consider the place of metacognition in the alternative assessments currently being promoted in the educational community. We conclude the chapter with discussion of general issues pertaining to the assessment of metacognition and recommendations for future directions.

DEFINITIONAL ISSUES

How metacognition is defined of course has important implications for how it is measured. The term initially was used by Flavell (1976) and by Brown (1978) in their early work in the 1970s to refer to knowledge about cognition and regulation of cognition. This two-component conceptualization of metacognition has been widely used in the literature since that time. However, Brown (1987) came to
believe that using the term to refer to two distinct areas of research creates confusion, clouding interpretation of research findings. In fact, White (1988) identified four possible facets to metacognition: (a) propositional knowledge about metacognition, (b) awareness of personal thinking, (c) ability to regulate thinking, and (d) readiness to apply that ability, and he wrote: “It is essential to know which of these are meant when an author refers to metacognition in order for communication to be clear” (p. 71). Some researchers have called for restricting its definition to knowledge about cognition (e.g., Cavanaugh & Perlmutter, 1982), excluding the regulatory processes. For example, Paris and his colleagues define the term as knowledge about cognitive states and abilities that can be shared (e.g., Paris, Jacobs, & Cross, 1987; Paris & Winograd, 1990). On the other hand, Sternberg (1991) believes that research on metacognition got off to a false start with its emphasis on what we know about our own thinking rather than on how we control our thinking.

Even today, there is still no consensus as to how metacognition should be defined. However, our own definition of metacognition includes both knowledge and control components (e.g., Baker, 1985b, 1994, 1996), and so we will be addressing measurement issues related to both. Those readers who prefer the more restrictive usage perhaps can be satisfied by thinking “cognitive monitoring” when we refer to metacognitive regulation or control.

Another definitional disagreement that has important implications with respect to measurement is whether metacognition is necessarily conscious. Some researchers have suggested that metacognition can be unconscious, tacit, and inaccessible (Pressley, Borkowski, & Schneider, 1987). However, the difficulty of measuring something that is unconscious and inaccessible is of course insurmountable, and therefore the position we have adopted is that metacognition refers to knowledge and control of cognition that is conscious or accessible to consciousness.

Two recent trends have expanded the scope of inquiry in metacognition, trends that other authors in this book have had a leading role in establishing. The first is the interest in “self-regulated learning,” which refers to learning that is self-directed, intrinsically motivated, and under the deliberate, strategic control of the learner (Pintrich & DeGroot, 1990; Schunk, 1989). The term self-regulation is sometimes used in the literature to refer to the use of skills included within the regulatory component of metacognition, such as planning, monitoring, and evaluating. For example, Borkowski, Day, Saenz, Dietmeyer, Estrada, and Groteluschen (1992) wrote that self-regulation is the “heart” of metacognition.
The second trend is the recognition that one cannot understand how and why people perform as they do on cognitive tasks without an examination of motivational and affective as well as metacognitive factors (Paris & Winograd, 1990; Pintrich & DeGroot, 1990; Pressley et al., 1987). Indeed, Borkowski, Pressley, and their colleagues (e.g., Borkowski, Carr, Rellinger, & Pressley, 1990; Borkowski et al., 1992) have argued that the “self-system” underlies the development of a metacognitive system. And Paris and Winograd suggested expanding the scope of metacognition to include affective and motivational aspects of thinking. In response to these new conceptualizations, measures of metacognition are often paired in research now with those that tap self-regulated learning as well as self-system factors such as attributional beliefs about the causes of success and failure and concepts of self as a learner.

METHODS FOR ASSESSING METACOGNITION USED IN BASIC RESEARCH

When one of us (LB) first set out in 1979 to synthesize the literature on metacognitive skills and reading for the Baker and Brown (1984) Handbook of Reading Research chapter, the term metacognition was seldom used. However, it was possible to identify a variety of methods that provided information about what we had defined as metacognition, even though it may not have been called this by the researchers who devised the measures. These methods are still widely used both in reading research and in other domains as well. To measure metacognitive knowledge about reading, researchers have relied on interviews and questionnaires. To measure metacognitive control in reading, or comprehension monitoring, researchers have used a variety of measures: detection of errors in passages; ratings of felt understanding; self-corrections during oral reading; completion of cloze tasks; on-line measures of processing during reading (e.g., eye movements and reading times); and retrospective or concurrent verbal reports (e.g., thinking aloud). In the chapter, we discussed the limitations of the various measures, and many publications since that time have also done so (Afflerbach & Johnston, 1984; Baker, 1985b, 1989; Garner, 1987; 1988; Pressley & Afflerbach, 1995; Winograd & Johnston, 1982). Because extensive discussions are available elsewhere, we will not devote much attention to these issues. However, because many of these measures are still in use in research and they are recommended for use by teachers and practitioners as well, it is important to summarize the relevant
issues here. We focus here on two approaches that are widely used but also widely criticized: verbal reports and the error detection paradigm.

Verbal Reports

One of the most frequently used approaches for assessing both metacognitive knowledge and metacognitive control is to ask students directly about what they know or what they do. Such self-reports have been collected in a variety of ways. For assessing metacognitive control, participants may be asked to think aloud about what they were doing and thinking as they solved a problem or read a text or to provide written comments periodically throughout the session (e.g., Bereiter & Bird, 1985; Cerro & Baker, 1993; Garner & Alexander, 1982). Or they may be asked to complete checklists of strategies they use (e.g., Phifer & Glover, 1982) or they may complete questionnaires or study strategy inventories (Cerro, 1995; Pintrich & DeGroot, 1990; Schraw & Dennison, 1994; Weinstein, Zimmerman, & Palmer, 1988). Students may be asked to report their strategies retrospectively or introspectively (e.g., Fischer & Mandl, 1984; Garner, 1982; Lundeberg, 1987; Winser, 1988).

Whereas verbal reports are but one way for assessing metacognitive control, they are the primary basis for collecting information about metacognitive knowledge, either through interviews or questionnaires (e.g., Belmont & Borkowski, 1988; Jacobs & Paris, 1987). In fact, many of the studies that assess metacognitive knowledge within a particular domain use questions that can be traced back to a few key studies. For example, most assessments of metamemory use at least some of the items used in the seminal study of Kreutzer, Leonard, and Flavell (1975). And many interview studies of children’s metacognitive knowledge about reading use questions from Myers and Paris (1978), which in turn were based on Kreutzer et al.

Research has convincingly shown that verbal reports of all types are subject to many constraints and limitations (Afflerbach & Johnston, 1984; Baker & Brown, 1984; Ericsson & Simon, 1984/93; Garner, 1988; Pressley & Afflerbach, 1995; Ward & Traweek, 1993). Briefly, problems with interviews include the following:

1. Participants may not be able or willing to express their thoughts and experiences.
2. Questions may not be understood by all participants.
3. Questions may induce responses based on social desirability.
4. Open-ended responses are often difficult to score.
Concurrent verbal reports (think-alouds) are also subject to many limitations, including the following:

1. Think-aloud procedures may disrupt processing of the task.
2. Cognitive processes may not be accessible to consciousness for report.
3. Personal characteristics such as age, motivation, anxiety, verbal ability, and willingness to reveal oneself may influence responding.
4. The instructions, types of questions, and probes that are used can cue participants to give particular kinds of responses.
5. The task needs to be difficult, complex, and novel enough to require metacognitive skills to perform.
6. Think-aloud protocols are difficult to score. (The coding scheme summarized in Pressley’s chapter [this volume] and described in detail in Pressley and Afflerbach [1995] is a welcome addition).

Despite their limitations, there is a general consensus that verbal reports can be valid and reliable sources of information about cognitive processes when elicited and interpreted according to guidelines recommended by such authors as Ericsson & Simon (1984/93). Advocates of this approach are sometimes impassioned in its defense. For example, Winser (1988) argued that self-reports are valid evidence of students’ processing, “in sharp contrast to the so-called objective and valid evidence from outmoded psychometric tests” (p. 260).

Error-Detection Approaches

The error detection paradigm is the most commonly used approach to assess metacognitive control in reading, that is, comprehension monitoring. It has also been used in listening situations (Baker, 1984; Flavell, Speer, Green, & August, 1981) and in research on mathematical problem solving (Van Haneghan & Baker, 1989). As used in reading, the reader is presented with texts that contain embedded problems or errors and is asked to identify them. The assumption underlying this paradigm is that these problems disrupt comprehension, and so the reader who is checking his or her ongoing comprehension should notice them. Much of the research in this area has shown that neither children nor adults are very successful at identifying the embedded problems (see Baker, 1985b, 1989; Baker & Brown, 1984, for reviews). Various measures have been used to
determine if readers are capable of detecting the errors: performance measures, such as underlining errors when they are encountered; verbal reports collected during or after reading; and on-line measures such as patterns of eye movements, reading times, and look backs (Baker & Anderson, 1982; Grabe, Antes, Thorson, & Hahn, 1987; Zabrucky & Ratner, 1992). However, caution is necessary in interpreting results of studies using this paradigm, as first discovered by Baker (1979) in her inaugural investigation of comprehension monitoring in adult readers.

In that study, students were instructed to read carefully six expository passages containing different types of embedded problems (internal inconsistencies, inappropriate logical connectives, and ambiguous referents) in preparation for answering subsequent discussion questions. After reading and answering questions calling for recall of the problematic sections of text, students were informed that the passages contained problems and were asked to report them, rereading as necessary. The students were also questioned as to whether or not they noticed the problems during reading, how they had interpreted them, and how they affected their overall understanding. Most surprising was that only 38% of the problems were detected, and fewer than 25% of these were reported to have been noticed during reading. Nevertheless, the recall protocols and retrospective reports made it clear that many failures to report problems were not due to failures to evaluate comprehension, but rather to the use of fix-up strategies for resolving comprehension difficulties. In other words, participants attempted to evaluate and regulate their comprehension, using strategies such as backtracking and seeking clarification in subsequent text. Thus, the study revealed the great lengths to which skilled readers go to make sense of text, especially if they have no reason to suspect that the texts were altered to be difficult to understand. Many studies conducted since that time have documented similar behaviors among elementary school children (e.g., Baker, 1984). There are clear differences in apparent comprehension monitoring effectiveness depending on whether readers are informed or uninformed about the presence of problems (e.g., Baker, 1984, 1985a; Baker & Anderson, 1982).

The 1979 study also revealed that adult readers use a variety of different criteria for evaluating their understanding; in fact, the participants frequently reported problems other than those intended to be conveyed. This led to the conclusion that failure to notice a particular type of problem embedded in a text does not necessarily imply poor comprehension monitoring (Baker, 1984, 1985a). For ex-
ample, the reader who fails to notice a contradiction within a passage presumably was not evaluating his or her understanding with respect to an internal consistency standard; however, he or she may have been using alternative criteria for evaluating comprehension. In much of Baker’s own research on comprehension monitoring, she has focused on the kinds of standards readers use to evaluate their understanding (Baker, 1985b), and has found that some standards are more likely to be applied than others, both by children and adults. What this means from the standpoint of measurement using the error detection paradigm is that care must be taken to specify exactly what aspects of comprehension monitoring one is interested in assessing and select embedded errors accordingly. Moreover, the information provided to participants is also critical; readers are more likely to identify problems when they know exactly what kind of problems to expect (Baker, 1985a; Baker & Zimlin, 1989).

Given the limitations of verbal reports noted earlier, exclusive reliance on post-reading verbal reports as a measure of error detection is unwise. Having participants underline problematic segments of text as they encounter them provides some evidence of online comprehension monitoring, but this performance measurement can only be used when readers are informed in advance of the existence of problems. With the increasing availability of affordable computers and appropriate software (Nason & Zabrucky, 1988), collecting process measures of comprehension monitoring while reading is becoming easier and more common. These measures include reading times and patterns of movement through the text (e.g., looking back, jumping ahead), measured either with eye movements or keystrokes.

Assessment of comprehension monitoring with the error detection paradigm is further complicated by demand characteristics of the task. Performance measures and verbal reports often give less indication of problem awareness than the online measures; the same reader who slows down when encountering inconsistent information may not report having noticed anything wrong (e.g., Harris, Kruithof, Terwogt, & Visser, 1981; Zabrucky & Ratner, 1992). Whether or not a problem will actually be reported depends on several factors: the participants’ goals for reading, the criteria they adopt for evaluating their understanding, and their threshold for deciding when a problem is serious enough to report. Moreover, personal characteristics play a role, such as whether an individual tends to be reflective or impulsive (Erickson, Stahl, & Rinehart, 1985); these findings lend weight to the importance of assessing the self-system concurrently with metacognition (Borkowski et al., 1992).
The error detection approach is often criticized for its lack of ecological validity on the grounds that typical texts do not contain embedded problems, but in fact this is not altogether true. We have been able to find “errors” corresponding to each of seven different standards of evaluation (Baker, 1985b) in naturally occurring prose. In other words, texts are often “inconsiderate,” and copy editors do not always do their jobs as well as they should. Zabrucky (1990) similarly argued that the paradigm is relevant outside the laboratory because of the prevalence of coherence problems in text. Nevertheless, such problems are not so prevalent that we can easily find suitable natural texts for our research. The reason researchers went to contrived texts in the first place is because skilled readers process text quickly and effortlessly when comprehension is proceeding well; it is only when obstacles arise that the process becomes slower and more deliberate. To increase the likelihood that obstacles would arise, embedded problems were deliberately introduced.

Despite the limitations of the error detection paradigm established through the research in the 1980s, a large number of studies continue to be conducted and published using the method. Unfortunately, many of them do not even take into account the cautions raised above. We feel it is time that we move beyond this approach in basic research on comprehension monitoring. It was a useful paradigm for providing insights into comprehension monitoring when research in that domain was in its infancy, and we have learned what we need to know from it.

Concerns Expressed about the Measurement of Metacognition

Virtually every empirical or theoretical article about metacognition includes at least an acknowledgement of the problems of measurement. In many cases, this acknowledgement is tied in with definitional issues: “The construct of metacognition and its measurement have remained somewhat elusive” (McLain, Gridley, & McIntosh, 1991; p. 84). Theory and research are impeded by difficulties that have been encountered in defining and measuring metacognition. In part, the problem has arisen because of the diversity of forms of investigation; there are few parallel studies or replications by independent researchers. Indeed, there are almost as many approaches to measuring metacognition as there are empirical research studies. This lack of consistency has occurred, in part, because the term metacognition has been used in many ways to refer to a wide variety of behaviors (Jacobs & Paris, 1987). Though such diversity is good in the early stages of
research on a topic, White (1988) noted, "eventually some sorting out is necessary" (p. 70) and we may now be at that point. Jacobs and Paris (1987) expressed a similar sentiment: "Now that the first glow of metacognition as a 'new approach' to reading has faded, the challenge is to continue to tackle the tough issues of defining, measuring, and fostering students' metacognitive approaches to reading" (p. 275). Other recent calls for more research on the measurement of metacognition have been made by Duffy et al. (1987), Wittrock (1991), Weinstein and Meyer (1991), Torgesen (1994), and Meltzer (1994).

As discussed earlier, many researchers, including ourselves, define metacognition as entailing both knowledge and control of cognition. Others, such as Paris and his colleagues, believe only the knowledge component should be subsumed under the label, thereby permitting direct measurement of metacognitions (Paris, Jacobs, & Cross, 1987). A major reason for their insistence on restricting the definition is that measurement of metacognitive control depends on inferences, saying: "Although these inferences may be warranted on occasion, they run the risk of assuming that children understand more than they actually do about the variables that influence thinking." (Jacobs & Paris, 1987, p. 264). However, there are many other researchers who would say that process measures are more valuable than verbal reports, the common means by which knowledge is assessed, because of inherent limitations in such measures (e.g., Clements & Nastasi, 1987).

Despite the importance attributed to metacognition, and the acknowledgement of measurement problems, little research has been conducted to test the adequacy of the measurement procedures, a concern expressed by many (e.g., Geary, Klosterman, & Adrales, 1990; Hertzog, Hultsch, & Dixon, 1989; Jacobs & Paris, 1987; Kirby & Moore, 1987; Meichenbaum, Burland, Gruson, & Cameron, 1985; Torgesen, 1994; Ward & Traweek, 1993). Torgesen observed, quite accurately, that research on metacognition has focused more on intervention than assessment, with the result that not much has been done to develop standardized assessment procedures that can be used as part of a diagnostic battery in applied settings. Others have argued that more work is needed to establish the construct validity of metacognition (Geary et al., Hertzog et al., Torgesen). Few standardized measures exist and many of those that do are not theoretically motivated (Meichenbaum et al.).

Many studies of metacognition and its relation to cognition, in both basic and applied settings, have yielded inconsistent results (Baker, 1994). Jacobs and Paris (1987) suggested that inconsistent
intervention outcomes may be due to metacognition being measured in different ways in different studies. And Schneider, Korkel, and Weinert (1987) suggested that failures to find strong correlations between memory behavior and metamemory may be due to the use of unreliable metamemory assessments. Rushton, Brainerd, and Pressley (1983) suggested that these weak relations may also have been due to the use of but a few items to measure metamemory, with resulting low reliability.

In response to criticisms such as these, some researchers have sought to develop standardized instruments that are theoretically motivated and that meet psychometric criteria of reliability and validity. In the next section, we consider some of these instruments, giving particular attention to those that have been subjected to independent testing by other researchers. We selectively discuss instruments in the following areas: metamemory assessment, metacognitive knowledge about reading, learning and study strategies (self-regulated learning), and problem solving.

DEVELOPMENT OF METACOGNITIVE ASSESSMENT TOOLS

Instruments for Assessing Metamemory

Research in metamemory has the longest history of any of the domains of metacognition, and some of the most stringent psychometric testing of instruments for assessing aspects of metacognition has been done in this area. We focus here on work done to develop and validate an instrument for assessing children's metamemory, undertaken by one of the other presenters at the symposium, John Borkowski. The instrument he developed along with several of his students and colleagues evolved from the classic metamemory interview of Kreuter, Leonard, and Flavell (1975). It consists of five subtests, three of which involve verbal reports alone and two of which involve metamemorial processing. The instrument initially was used with second graders as an individually administered test by Kurtz, Reid, Borkowski, and Cavanaugh (1982). Reliability and validity were considered adequate; test-retest correlations for subtests ranged from .29 to .49, though the composite was considerably higher, .67. The metamemory battery was later adapted for group administration by Belmont and Borkowski (1988) and was tested with third and fifth graders. Age-related differences were found on each of the five subtests, consistent with theoretical predictions. Correlations among the subtests were near 0, suggesting metamemory is task- or domain-specific rather than general. The overall test-retest reliability of the Metamemory
Assessment Battery was very similar to that of the individually administered instrument, .66. The group-administered battery was independently tested for validity by Geary, Klosterman, and Adrales (1990). Geary et al. looked for age-related changes among second and fourth graders as one way of establishing validity; they found age-related differences on all but one subtest and evidence that the test might be too difficult in general for second graders. Geary et al. also found correlations near 0 among the subtests. There were some significant correlations with achievement test performance, providing some evidence of convergent validity. The authors concluded that their study provided some converging evidence for the validity and utility of the battery, but that more information was needed as to appropriate age ranges.

Instruments Designed to Assess Metacognitive Knowledge in Reading

As noted earlier, there have been numerous studies of metacognitive knowledge about reading involving both children and adults. Most of these studies have used structured interviews with open-ended questions. Few efforts have been made to develop interview instruments intended for use beyond the research setting of the study, with perhaps the only exception the work of Kirby and Moore (1987). Nevertheless, as we will see in a subsequent section, interviews are widely recommended for use in classrooms and clinics.

We focus here on a multiple-choice questionnaire, the Index of Reading Awareness (IRA), developed originally as a research tool and recommended for use to classroom teachers as an informal assessment instrument (Jacobs & Paris, 1987; Paris, 1991). The goal was to design a measure that would be sensitive to individual and age-related differences in awareness about reading and to changes in awareness occurring during a school year and/or in response to instruction. According to Jacobs and Paris, the IRA assessed "children’s knowledge about reading and their abilities to evaluate tasks, goals, and personal skills; to plan ahead for specific purposes, to monitor progress while reading, and to recruit fix-up strategies as needed" (p. 268). The IRA assessed planning, evaluation, and regulation, using 15 items from the Paris and Jacobs (1984) interview, with three response options based on children’s actual answers given to the interview items. Another five questions assessed knowledge about strategy utility, the understanding of when and why particular strategies should be used. Choices are awarded 0, 1, or 2 points, corresponding to inappropriate, partially adequate, or strategic responses. The IRA was designed for third to fifth graders, with grade equivalent reading abilities in second through seventh grade. Its
use in a study evaluating the effects of an intervention that incorporated metacognitive instruction showed the instrument was sensitive to changes in awareness due to individual differences in age, sex, and reading ability.

The IRA was subjected to an independent test of reliability and validity by McLain, Gridley, and McIntosh (1991), who felt that the psychometric properties of the instrument had not been adequately tested by Jacobs and Paris (1987). McLain et al. administered the IRA to 145 children in grades 3, 4, and 5. Tests of reliability revealed that the Cronbach’s alphas for the four subscales were low (.15–.32), and the total reliability for the items was .61 and for the subscales .56. A preliminary factor analysis did not yield interpretable factors. McLain et al. also tested validity by determining whether the instrument yielded the theoretically predicted age-related increases in awareness. They did find such differences between third and fifth graders, as did Jacobs and Paris, but fourth graders were comparable to fifth graders in the McLain et al. study, leading the authors to conclude that “conceptualizing metacognitive awareness as increasing steadily with age may be erroneous” (p. 86). Tests for criterion-related validity revealed that although the IRA was moderately correlated with standardized reading comprehension scores, once basic reading skills were controlled for statistically, the IRA added little or no information to the prediction of comprehension.

McLain et al. (1991) concluded that the IRA “should be used cautiously as a measure of metacognition in reading for both research and classroom use” (p. 86). Their analyses questioned both the internal and criterion-related validity of the scale. They considered the scale to be acceptable “if used as a total score and only as one measure of the reading process in a portfolio assessment” (p. 86). Moreover, the subscale scores should not be used separately because internal consistency reliability was too low. Paris (1991) himself wrote that separate scores should not be reported because the four constructs tapped by the scale are not independent.

In their description of the development of the IRA, Jacobs and Paris (1987) argued that the multiple-choice format avoids some of the pitfalls of verbal reports. Specifically, it is more objective than interviews that may involve interpretations of open-ended responses, experimenter bias, or fabricated responses; it does not put shy or inarticulate children at a disadvantage; the measure is based on empirical research of children’s responses to metacognitive questions; it accurately reflects children’s knowledge about reading strategies rather than researchers’ beliefs about what children know; and it is easier to administer, in that it can be given to groups rather than
individuals, it can be completed in a short time, and it is easy to score. However, there are dangers associated with such a format. Duffy et al. (1987) evaluated their own efforts to develop a multiple-choice instrument to assess students’ awareness of strategy use in reading and identified problems that are relevant to all attempts to develop multiple-choice assessments of metacognition. One is that the multiple-choice format suggests there is a single right way to think about using a particular strategy, a criticism also made by Rhodes and Shanklin (1993) in a critique of metacognitive instruments advanced for use in the classroom, such as that of Schmitt (1990). Another concern is that it is difficult to write distractors that are plausible. Those used in the Duffy et al. study were, however, considerably less plausible than those used in the IRA, which avoided this problem by using only options provided by children during earlier interviews. Another problem with multiple-choice assessments such as the IRA is that they “could easily be corrupted by teaching children to mimic stock answers to the questions” (Paris, 1991, p. 38). For this reason, Paris argued, the IRA was not intended to be a formal assessment of metacognition, but rather, should be used only informally.

Instruments for Assessing Metacognitive Strategy Use in Learning and Studying

Instruments that include assessments of metacognitive functioning in learning and study situations had their origins in early inventories of “study skills” that tended to focus on overt behaviors such as underlining and note-taking. Many instruments designed for this purpose have been developed in recent years, most of which are intended for use by adolescents and adults (college students). We will briefly note a few relevant instruments.

The instrument developed by Pintrich and his colleagues, the Motivated Strategies for Learning Questionnaire (MSLQ), is a self-report measure of adolescents’ and college students’ motivational orientations and use of various learning strategies (Pintrich & DeGroot, 1990; Pintrich, Smith, Garcia, & McKeachie, 1993). Thirty-one of the items are motivational and 50 are learning strategies. Of these 50, 12 items are concerned with metacognitive self-regulation; specifically tapping the processes of planning, monitoring, and regulating. The instrument has been subjected to extensive checking of reliability and validity. The metacognitive self-regulation scale has an adequate level of internal consistency (.79), and an overall correlation of .30 with final course grade.
Perhaps the most widely used instrument is the Learning and Study Strategies Inventory (LASSI, Weinstein, 1987), available as a paper-and-pencil test and in a computerized version. The LASSI consists of 77 items rated using a 5-point Likert-type scale that ranges from not at all typical of me (1) to very much typical of me (5). It yields 10 subscale scores: attitude, motivation, time management, anxiety, concentration, information processing, selecting main ideas, study aids, self-testing, and test strategies. The assignment of items to these subscales was based on the intuitive judgment of several experts (Weinstein, Zimmerman, & Palmer, 1988). The LASSI has been subjected to extensive validation efforts, including those by independent researchers (e.g., Olejnik & Nist, 1989), and it is considered to have good psychometric properties.

A promising new instrument focusing more exclusively on metacognitive awareness was developed by Schraw and Dennison (1994). The inventory consists of 52 items in which the respondent indicates how true the statement is of him or her on a 100 mm. scale. Some of the items tap an individual’s knowledge about cognition (declarative, procedural, and conditional knowledge) and some tap regulation of cognition (planning, information management strategies, monitoring, debugging strategies, and evaluation of learning). Factor analyses revealed these two factors had good internal consistency (.90) and were intercorrelated (.54). The authors interpret their results as providing support for the two-component conceptualization of metacognition; however, their focus really is on two types of knowledge, rather than knowledge and regulation per se.

Instruments for Assessing Metacognition in Problem Solving

All of the other instruments discussed thus far, with the exception of portions of the metamemory assessments, rely on self-reports of metacognitive knowledge or control. Within the domain of problem solving, there are self-report instruments as well as process measures that provide on-line evidence of metacognitive control.

Several assessment approaches, both process-oriented and self-report, have been based on Sternberg’s (1986) metacomponential theory. In Sternberg’s theory, metacomponents are the metacognitive or executive processes used in planning and evaluating cognitive activities. Sternberg has developed paper and pencil measures for research purposes, where the use of metacomponents is inferred on the basis of response time and accuracy, but he does not yet have an instrument he recommends for formal assessment (Sternberg, 1991).
Clements and Nastasi (1987) developed a naturalistic approach to measuring metacomponential processing, arguing that a naturalistic setting was needed to increase ecological validity. In the study reported for instrument development and validation, children worked together in pairs to solve various kinds of problems, and all verbalizations were coded as to the types of metacomponential processing involved. The authors concluded that both reliability and construct validity of the observational instrument were acceptable; interrater agreement of the classification of metacomponential processes was 87%, and there were significant correlations between the observational task and paper-and-pencil tasks. Clements and Nastasi discussed their approach as an instrument with practical utility (the article was published in *Psychology in the Schools*), but it cannot really be picked up easily and used in educational settings because it requires careful analysis of verbal protocols. Moreover, it is difficult to disentangle the relative contributions of the two children who are observed. The approach warrants further research and refinement before it should be recommended for use by educators. Swanson (1990) also assessed component processes of problem solving (using his own system based on analysis of think-aloud protocols), and he found that these process measures were related to verbal reports on a metacognitive interview (modeled on Kreutzer et al., 1975) focused on problem solving.

Instruments designed as self-report measures of metacomponential processing have been developed by Armour-Thomas and her colleagues. The Student Thinking About Problem Solving Scale (STAPPS) consists of 37 items and has been subjected to two separate factor analyses, which yielded markedly different results, even though the populations were similar. Armour-Thomas and Haynes (1988) administered the STAPPS to high school students (predominantly African American and Hispanic) and obtained a six-factor solution accounting for 73% of the variance. In contrast, Armour-Thomas, Bruno, and Allen’s (1992) factor analysis yielded three different factors which accounted for 29% of the variance. The inconsistencies in the results of the factor analyses are of course problematic and reveal that this instrument is not ready for general use. Perhaps in recognition of this problem, Allen and Armour-Thomas (1993) developed another self-report instrument of metacomponential processing, with items tapping use of each of Sternberg’s eight metacomponents in four different domains, both academic and nonacademic. The theoretical underpinnings of the instrument are solid, but once again the validation efforts were less than satisfactory, with factor analysis yielding what to us appear to be uninterpretable results. This may
well be due to problems with the items themselves, responses to which are likely influenced by social desirability factors.

Meltzer (1991, 1994) has developed an instrument intended to be multidimensional, tapping metacognitive and strategic processing in several different domains. We include it in this section because of its emphasis on problem solving. The Surveys of Problem Solving and Educational Skills (SPES) "represents one of the first pilot attempts to systematize some of the informal approaches used currently in clinical assessment for the evaluation of students' metacognitive awareness and reliance on strategic learning" (Meltzer, 1994, p. 598). Unlike most of the other instruments discussed, which had their origins in basic research on an aspect of metacognition, the SPES was specifically designed for diagnostic use in clinical and school settings with children aged 9 to 15 with learning difficulties. Meltzer argued that there is a need for procedures that evaluate metacognitive strategies as they interact with cognitive processes such as problem solving, language, memory, and attention. The SPES actually consists of two separate parts: The Survey of Educational Skills measures strategic performance in the academic areas of reading, spelling, written language, and mathematics. The Survey of Problem Solving measures strategic problem solving on six different tasks, three nonlinguistic and three linguistic/verbal. The SPES is based on a model that focuses on major features of strategy selection that are essential for learning: efficiency, flexibility, methods, styles (self-monitoring, systematic and planful, reflective), and the ability to justify the solutions provided. It emphasizes the importance of systematic observations of the learning strategies and processes used by students in different situations. Response demands include think-alouds, retrospective reports, and introspection on strategies used. Systematic observations of how the student approaches the tasks and analyses of error patterns are also important features of the assessment.

The SPES holds great promise as a process-oriented assessment tapping important aspects of metacognitive control. However, Torgesen (1994) expressed concern that if the SPES came to be used widely in diagnostic work, "it might create the impression that metacognitive processes can be usefully measured and perhaps remediated, as a set of domain-general skills" (p. 156). His concern stems from doubts about the domain generality of metacognition (this issue is discussed in more detail subsequently). Nevertheless, Meltzer (1994) herself did stress that the SPES should not be used as a method for analyzing and then training domain-general problem-solving processes.
Very early on in the history of metacognitive research, recommendations began to appear in the literature for teachers, summarizing the research findings, emphasizing their educational significance, and suggesting ways for teachers to promote metacognition in the classroom and to assess it informally in their students. During our literature search for preparation of this chapter, we found that many of the articles that addressed metacognitive assessment appeared in journals for teachers and practitioners such as school psychologists. This is consistent with the finding by Paris, Wasik, and Van der Westhuizen (1988) in their literature search covering the years 1981-1987. Of the 124 journal articles they found, only 40 were empirical; the rest “extoll(ed) the virtues of metacognition for understanding reading” (p. 163). They argued, as have we in similar terms (Baker, 1994), that there is a “dangerous imbalance in which the enthusiasm and prescriptions far outstrip the empirical data base” (p. 163). Many of the recommendations we found appear to be based on limited empirical evidence. Several articles and books include actual instruments that teachers can use, but most of these instruments have little or no validation. In this section we consider the prescriptive advice given to teachers and school psychologists for how they might assess metacognition using interviews, think-alouds, error detection, and process measures.

Recommendations for Using Interviews

Almost every article written for teachers or practitioners about metacognition includes recommendations to interview students about their metacognitive knowledge and strategy use. However, a lack of explicit information as to how to use the interview information, and a lack of a caution on the limitations of verbal reports, is typical of a number of these articles. For example, Ellis (1989) included sample questions for teachers to use in a metacognitive interview, but he did not provide any guidance as to how teachers should use the information or interpret the students’ responses. He simply wrote that the interview’s purpose is to find out what students know about their own thinking, their perceptions of their own thought processes and cognitive strategies, and their perceptions of strategies they were asked to use. Garner (1992) suggested teachers can interview readers to get a sense of their views of the reading process and their knowledge of reading and study strategies using questions originally designed for research purposes. However, she was careful to caution
teachers of the need to be aware of the limitations of verbal reports if they interview their students, explaining problems of accessibility, memory failure, inadvertent cuing, and verbal facility.

Several different authors have recommended the use of either the interview questions originally used by Paris and his colleagues (e.g., Myers & Paris, 1978; Paris & Jacobs, 1984), their multiple-choice Index of Reading Awareness (Jacobs & Paris, 1987), or both. These include Paris (1991), Zabrucky and Ratner (1990), and Lloyd and Loper (1986). Lloyd and Loper recommended for their school psychologist audience that they begin by determining if students can respond to the IRA questions open-endedly; if not, then the multiple-choice options should be provided.

Many of the recommended interviews include items that focus on students’ views of themselves as readers (e.g., What do you do best when you read?), consistent with the recent focus on self-system factors. For example, Yochum and Miller (1990) stressed the importance of considering both metacognition and attributions and achievement motivation. Others who have recommended interviews include Gray (1987), Weinstein and MacDonald (1986), and Paratore and Indrisano (1987).

A number of published interview instruments that have been recommended for teachers are now being published in secondary sources, thus giving them what might appear to be even greater legitimacy. For example, Rhodes (1993) published a handbook of informal instruments for assessing literacy that included several metacognitive interviews drawn from other sources. One of the instruments was the 10-question Reading Interview: A reader’s view of the reading process (Goodman, Watson, & Burke, 1987). Directions for administering the interview provided by Goodman et al. include coding directions, with categories of responses students might provide. This level of detail seems appropriate and helpful for teachers. Another interview Rhodes included in her collection was a content reading interview based on Wixson, Bosky, Yochum, and Alvermann (1984). The questions are similar to those used in other interviews, but are tied specifically to a particular content area selected by the interviewer. This instrument has been criticized on the grounds that no reliability data or validity data were provided, but the lack of such data is a common weakness of most of these interview instruments.

Recommendations for using think-aloud measures

The growing popularity of think-aloud procedures in research on cognitive processing and metacognition has led, not surprisingly, to recommendations for its use as a diagnostic tool. As with the inter-
There is often a lack of explicit attention to the problems inherent in collecting think-aloud protocols and the ways that the data should be interpreted. Most of the recommendations have been addressed to those who work with college students as opposed to younger children, perhaps because there is still uncertainty as to how effectively children can engage in productive think-alouds. For example, Randall, Fairbanks, and Kennedy (1986), Nist and Kirby (1986), and Steinberg, Bohning, and Chowning (1991) advocated using think-aloud procedures with college students experiencing reading difficulties. Steinberg et al. explicitly acknowledged that the complex coding systems used in research analyses of think-aloud protocols would not be appropriate for teachers to use but they did not offer simpler alternatives. Winser (1988) recommended using think-alouds with students of all ages and abilities, including children as young as second grade. Yochum and Miller (1990) also recommended collection of think-alouds with elementary-aged children. Winser reported working with several teachers who confirmed that the think-aloud approach could be used for evaluation.

Think-aloud approaches have also been recommended as informal assessments in math as well as in reading. For example, Lawson and Rice (1987), in an article written for school psychologists, discussed the value of having students think aloud as they solve math problems. This would help the teacher diagnose difficulties the student has with respect to problem solving and allow for analysis of error patterns. The authors included a simple-to-use “coding schedule” that includes items such as metacognitive knowledge that is made explicit, checking, planning, and strategy use.

Recommendations for Using Error Detection Procedures

Several investigators who have conducted research using the error detection paradigm and have identified problems with it in their empirical reports have gone on to write articles for teachers recommending its use in assessment. Although some caveats are included, they do not seem strong enough to us. For example, Garner (1992), Zabrucky and Ratner (1990) and Paris (1991) have all recommended this approach for assessing comprehension monitoring. Zabrucky and Ratner (1990) wrote that the ability to evaluate comprehension “is assessed by introducing errors into passages,” implying that this is the only way possible. They recommended adapting grade-appropriate texts, introducing different kinds of problems to find out what standards children can use and what standards they
need help using. The authors cautioned about reliance on verbal reports of error detection, and asserted that underlining is a better indicator of what children can do than are answers to questions. However, to our knowledge, this assertion is not supported by empirical evidence. To assess the ability to evaluate, Zabrucky and Ratner advised, give children specific information about the nature of the problems and examples; to assess spontaneous evaluation, they continued, do not forewarn children that passages have problems. This latter recommendation seems problematic to us because children may well be spontaneously evaluating using criteria other than those represented by the embedded problems.

Garner (1992) identified some of the difficulties researchers have had in disentangling explanations for poor detection performance, but encouraged teachers “to experiment with error-detection exercises in the classroom” (p. 244). She suggested teachers could assess children’s use of different standards of evaluation through the process of embedding errors in short expository passages, asking children to underline anything troublesome, and having them explain the nature of the problem. Garner reported that teachers she has worked with found this procedure useful in revealing whether there was reliance on one particular type of standard. Garner offered the good advice that work with contrived texts should be phased out to work with uncontrived texts.

Paris (1991) also recommended the error detection approach, saying that it can be adapted easily for diagnostic and remedial purposes. He described various kinds of errors that can be introduced. He listed the following advantages of the approach: It can be used with regular curriculum materials and may be particularly useful in content area reading; it can be used with individuals, small groups, or large classes; and it can be used as a paper-and-pencil silent reading task or it can be given orally. “Besides the flexibility, quick administration, adaptability to the reading level of each student, and the savings in time and money with a locally designed task, error detection tasks promote a thoughtful, inquisitive interaction while reading, so that the goals of instruction and assessment are congruent” (p. 39).

Others who have recommended error detection procedures include Gray (1987), who did explain for teachers why failures to notice errors may not signal poor comprehension monitoring, and Weinsten and MacDonald (1986), writing for school psychologists without critical commentary on the approach.

We have been rather critical of these recommendations for using the error detection paradigm to assess children’s ability to monitor
their comprehension. However, we too, have written about using error detection methods in the classroom (Baker, 1991), and we have incorporated the method in a metacognitively oriented curriculum for customer service workers as part of a workplace literacy program (Baker et al., 1994). We think it is a useful instructional tool for helping readers to see the variety of ways that comprehension can fail and the variety of things that can make text difficult to understand. But we do not believe it should be used for formal assessment purposes. Use for informal assessment is perhaps acceptable if the tester is well aware of its limitations and it is used in conjunction with other assessment approaches. But it should not be used in group-administered paper-and-pencil assessments because the risks of misinterpreting failures to detect problems are too great. In group administration, students are typically presented with passages containing problems and asked either to underline problems, to write down what if anything did not make sense, or to rate how well they understood the passage. Without the opportunity for an individual interview, we cannot be sure why a reader may not have identified the intended problems.

Recommendations for Assessing Metacognitive Processing in Authentic Tasks

The simplest recommended process assessment is to observe students while they are engaged in authentic tasks such as reading, writing, or mathematical problem solving. Zabrucky and Ratner (1990) advised that given the problems with verbal reports, teachers may need other approaches to assess what children do instead of what they say they do. They suggested that observing children while they read may provide the best assessment of regulation of comprehension, but they did not give specific guidance as to how to do this. Others who recommended naturalistic observations include Yochum and Miller (1990) and Lloyd and Loper (1986). Several books have been published that include observational checklists for use by teachers interested in assessing literacy, including Burke (1993); Kemp (1990, cited in Paris, 1991); Rhodes (1993); and Rhodes and Shanklin (1993). For example, Kemp included observational records that can provide information about strategies, metacognition, and motivation in authentic tasks.

One recommended approach that has a number of advantages is to collect “running records” to evaluate children’s oral reading strategies (Paris, 1991; Rhodes & Shanklin, 1993). As the child reads aloud, the teacher records oral reading miscues, including substitutions,
rereadings, omissions, and self-corrections. Winser (1988) recommended an interesting variation of this procedure involving stimulated recall: Children read a passage orally; the session is videotaped and the children are asked to talk about their self-corrections (e.g., “What did you do when you fixed that part up?”). The value of this sort of data, Winser asserted, is that it “provides teachers with some clues to the way their students are actually functioning, so that they have an insight into their learning styles that is not available from traditional tests” (p. 264). Retrospective analyses of running records have an advantage over traditional verbal reports based on hypothetical or “typical” behaviors in that they focus the individual’s attention on a particular task context. However, the time-intensive nature of this procedure may make it more suitable for research purposes than for practical assessments. Another advantage of approaches involving running records, which also applies to think-aloud procedures, is that they can be used with naturally occurring materials and so have greater ecological validity than error detection procedures.

Another authentic approach was developed by Paris (1991) for assessing children’s reading comprehension as well as their strategies, motivation, and metacognition. The “think-along” approach, recommended to teachers and clinicians, simulates a real classroom experience where the student reads aloud and the teacher asks interspersed questions. The questions not only assess understanding, but also how students know they know the answers, or if they do not know, how they can find out. The teachers probe students’ thinking with questions about their strategies and also observe spontaneous strategy use. The approach is available commercially as the Heath Reading Strategies Assessment (1991), but Paris stressed that any passage can be used as a think-along passage. He included in his article generic questions that can be used to assess both comprehension and metacognition. The students’ responses are evaluated with respect to strategy effectiveness, but the burden of judgment is on the examiner or teacher, as it is in most of the recommended approaches. An answer sheet has spaces for checking off the strategies used for identifying the topic, predicting, monitoring meaning, making inferences, and summarizing. For example, the teacher might question the child about an unfamiliar word: “What do you think ‘trat’ means in the sentence you just read? How could you tell? If you don’t know, how could you find out?” The checklist of strategies includes: uses context cues, substitution looks or sounds similar, mentions others as resources, and mentions dictionary as resource.
Garner (1992) suggested still another approach in which metacognitive knowledge could be revealed in an authentic setting: observation of peer tutoring. One child serves as tutor for another; the tutor is the focus of particular interest in this assessment. Listening to how the tutor describes strategies to a child who is not using them spontaneously provides insight into the tutor's own metacognitive knowledge. Does the tutor show awareness, for example, of how to use reinspection to locate information in a text that the tutee could not remember?

Comments on the Recommendations for Metacognitive Assessments in Classroom and Clinic

As the preceding review should make clear, there have been many recommendations for teachers and school psychologists to assess metacognition, dating back at least as far as the mid 1980s (e.g., Bondy, 1984; Weinstein & MacDonald, 1986). The literature for practitioners extends to school counselors as well; Mills and Brunner (1988) wrote about the need for school counselors to be aware of metacognition and of ways to assess it in their clients (students). As should also be clear, we have serious reservations about the way many of these recommendations are framed. Those made to school psychologists are perhaps less problematic than those made to teachers. School psychologists have advanced degrees that involve training in assessment techniques, and they should also be better prepared to be critical consumers of the literature. Classroom teachers, on the other hand, frequently have little formal training in either research methods or assessment, and so they are more likely to take the recommendations at face value. Researchers who write for teachers, who attempt to translate research into practice, have an ethical obligation to frame their recommendations responsibly, providing concrete advice on how to interpret the data that may be collected through interviews, think-alouds, and error detection tasks. The same is true to some extent for researchers writing for school psychologists, who may not have the time to familiarize themselves with the primary sources on which the recommendations are based. It is important that teachers not be left with the false impression that they can easily acquire useful or meaningful information by administering these measures.

Many of the materials written for teachers overgeneralize the construct of metacognition to refer to the use of any kind of strategy during cognitive activity, a practice that has led to some confusion in the literature and fueled recommendations to restrict the term to knowledge about cognition (e.g., Brown, 1987). For example, in a
book that consists of a collection of assessment instruments, Burke (1993) included a listing of the following “metacognitive” abilities to look for: “ability to solve problems and to make decisions; ability to brainstorm or generate ideas.” Further overgeneralization occurred in her recommendation for teachers to use journals as “metacognitive strategies” by assessing the reflectiveness of the student’s response.

Despite the plethora of recommendations, it is not clear how widely they have been adopted. Garner (1992) wrote about teacher assessment of metacognition as though it were commonplace: “Many teachers assess what their students know (and don’t know) about the reading process in general and about important reading and study strategies in particular” (p. 242). But is it? No data addressing this question are available to our knowledge, although it does appear that the emphasis on the importance of metacognition has reached the classroom teacher. Commeyras, Osborn, and Bruce (1993) studied teachers’ reactions to items on the 1992 National Assessment of Educational Progress (NAEP), which included a special study of fourth grade students designed to examine their awareness of their own comprehension. Their use of effective reading strategies was assessed, analyzed, and reported as descriptive data. Teachers were asked the extent to which they believed the study was needed. Responses were obtained from 312 teachers, 80% of them at the elementary level. Forty two percent gave the highest rating of 5 (to a very great extent); 36% the next highest rating of 4; 14% gave a rating of 3; 4% gave a rating of 2; and only 3% gave the lowest rating of 1 (not at all). Thus, the majority of teachers who responded to the survey appeared to believe this type of metacognitive assessment was important.

To what extent are metacognitive assessments used in diagnostic settings? Again, little information is available, but some relevant data were collected in England. Farrell, Dunning, and Foley (1989) conducted interviews in England with 100 school psychologists in 1981 and 1986 to determine the types of instruments used to assess children with learning difficulties. Their conclusion was that psychologists have hardly begun to assess children’s metacognitive strategies and that practice has only partially kept up to date with developments reported in the literature.

The Place of Metacognition in General Assessments of Educational and Intellectual Functioning

Traditional approaches to intellectual and educational assessment do not reflect metacognitive skills, and there is a growing
demand for change in this direction. With respect to intellectual assessment, Carr and Borkowski (1987) wrote, “The inclusion of process-oriented measures (e.g., metamemory and components of metacognition) in the assessment of intelligence may minimize the need for product-oriented measures which often fail to provide educationally valuable information about learning skills and deficiencies” (p. 43). Sternberg (1991) also believes that intelligence tests should put greater emphasis on metacognition, and the test he is developing based on his componential processing theory includes assessments of metacomponential processing (i.e., metacognition). With respect to educational assessment, Benton and Kiewra (1987) discussed the need for metacognitive assessment in the academic domains of reading, writing, and mathematics. And Glaser, Lesgold, and Lajoie (1987) identified metacognitive skills for learning as a dimension that should be assessed in the measurement of achievement. Many of the recommendations have as a premise the need to make assessment practices more in line with current views of learning and instruction. For example, the prevalent view of reading as a strategic activity has led to calls for reading assessment to incorporate metacognitive assessment (Duffy et al., 1987; Valencia & Pearson, 1986).

Critics of traditional tests argue that intelligence tests are insensitive to student’s metacognitive and attributional perceptions of the task, strategies, and personal abilities, and therefore these psychometric evaluations are not very relevant to educational intervention (Paris, Jacobs, & Cross, 1987). The focus on static levels of performance rather than on emerging cognitive processes provides little direction for intervention. Current educational achievement tests also are not very successful at diagnostic testing because they do not reveal the processes by which a response to a problem or question is constructed and so do not reveal the types of misunderstandings that individual students have (Linn, 1991). Accordingly, there are many calls for new modes of assessment that focus on the processes of cognitive activity rather than the products (e.g., Carr & Borkowski, 1987; Clements & Nastasi, 1987; Ellis, 1989; Linn, 1991; Paris et al., 1987; Mills & Brunner, 1988; Meltzer, 1994; Taylor, 1987; Ward & Traweek, 1993), and also for more “authentic” forms of assessment that capture what students do in more ecologically valid contexts. We now consider briefly the place of metacognition in some of these alternative assessments, including dynamic assessments, portfolio assessments, and performance assessments, both commercially available instruments and statewide performance assessment programs.
Dynamic Assessment

Dynamic assessment approaches are becoming increasingly popular as a way of assessing the processes of learning, including metacognitive control, rather than the products of learning that are assessed in traditional static measures (Ellis, 1989; Kaniel & Reichenberg, 1990; Lidz, 1991; Linn, 1991; Meltzer, 1994; Paris et al., 1987; Ward & Traweek, 1993; Taylor, 1987). These approaches, also known as mediated assessment, assisted learning, and learning potential assessment, view instruction and assessment as closely intertwined. The distinctive feature of dynamic assessment is that it includes a teaching phase. The students' independent performance is first assessed, followed by instruction and subsequent retesting. This test-teach-retest method allows the students' responses to intervention to be examined, revealing cognitive and metacognitive processes that are available but not necessarily used. The teaching phase can include instruction in both cognitive and metacognitive aspects of the task.

Recommendations for dynamic assessment as an alternative to traditional psychometric tests are appearing in the literature for teachers and practitioners. Ward and Traweek (1993) provided an illustration of how think-alouds could be used by school psychologists in dynamic assessment, addressing the question of whether students needed only a simple prompt to activate metacognitive awareness and strategic processing. Weinstein and MacDonald (1986) also recommended that school psychologists use a process approach to determine if students have learning problems because of cognitive monitoring deficits: Form hypotheses about the source of the problem, teach specific strategies, and assess whether the strategy has helped the child's performance.

Within the specific area of reading, there have also been similar recommendations. Ellis (1989) described a model for assessing students' use of reading strategies and their metcognitive knowledge about reading that included obtaining process measures of strategic functioning via mediated cues to use various cognitive strategies while reading. Paratore and Indrisano (1987) also proposed a mediated assessment of reading comprehension: First give comprehension tasks traditionally; if there are difficulties, initiate intervention with the instructor teaching the student a strategy and modeling its use; then administer a new passage and observe the student's use of the strategy.

Portfolio Assessments

Many educators have advocated the use of portfolios to capture real uses of literacy, math, or science. Just as artists create portfolio collec-
tions to display their best work, so too, it is argued, should students. Much has been written about portfolios in authentic assessments of literacy in particular (Valencia, Hiebert, & Afflerbach, 1994). Portfolio assessments involve metacognition because students' written reflections about themselves as learners and about their learning typically are critical components (Hansen, 1994; Snider, Lima, & DeVito, 1994; Valencia & Place, 1994). Having students keep daily "learning logs" (e.g., Bondy, 1984) also provides a means by which teachers can assess students' awareness of their own cognitive processes. However, Valencia and Place (1994) suggested that teachers should first provide modeling and guided practice in metacognitive reflection because this is not something many students do spontaneously.

Commercial Performance Assessments

Given the limited number of assessment instruments that have been documented as reliable and valid, it is not surprising that there are very few commercial instruments available. As Lloyd and Loper (1986) noted, because there are no norm-referenced commercial instruments for the assessment of metacognition, school psychologists must develop their own assessment procedures. There is apparent demand, however, for we are beginning to see some attention to metacognition among commercial test publishers (Linn, 1991; Paris, 1991; Powell, 1989). Paris (1991) discussed some of the instruments available in reading that include metacognitive assessments, such as: the Qualitative Reading Inventory (Leslie & Caldwell, 1990) and the Heath Reading Strategies Assessment (1991) that incorporates the "think-along" approach developed by Paris himself to assess comprehension and metacognition simultaneously.

Statewide Performance Assessments

The new statewide performance assessments that are being used in such states as Michigan, Illinois, and Maryland include measures of metacognition, in response to the growing awareness that assessments should include evaluation of thinking skills, strategy use, and metacognition. These assessments are designed for group assessment only, however; individual scores are not reported because not all students receive the same tasks and generalizability cannot be assured. The Michigan items measure children's knowledge about reading (e.g., the strategies that are appropriate for different purposes). The Illinois test poses scenarios to students and asks them to
judge whether particular strategies would be helpful or not in those circumstances. For example, students might be given a scenario in which they are asked to retell a selection they just read to different audiences: a peer, a younger child, and a teacher. Then they rate the helpfulness of several different responses for each audience (Valencia & Pearson, 1986). The Maryland State Performance Assessment Program (MSPAP) also examines metacognition; one of its outcomes is demonstrating awareness of strategic behaviors and knowledge about reading. This information is gathered through questions such as the following used in pilot work (Kapinus, Collier, & Kruglanski, 1994, p. 265):

> When you read a story such as the Great Kapok tree, you may come to a part that you don’t understand. Put a check mark in front of each thing below that tells what you might do. You may choose as many as you want. If you do something that is not listed, write it on the line next to the word “other.”

Sometimes I

- keep reading and then come back to that part
- skip over the part that is confusing
- ask someone about the part that is confusing
- try to sound out new words
- use a dictionary
- other: ____________________

There is a danger with test items such as this that students may respond correctly about the strategies they would use because they have been coached, but the knowledge would not transfer to authentic situations (Wixson, 1994). Recall Paris’ (1991) caution that the Index of Reading Awareness, which includes similar kinds of questions, should not be used as a formal assessment because of the danger of mimicking stock answers.

The Arizona Student Assessment Program (ASAP), as described by Garcia and Verville (1994), does not have metacognitive assessment as an explicit goal, unlike the other three state programs. However, it includes what we have called metacognitive control strategies in its comprehension outcomes: “uses strategies to self-correct when necessary,” with the associated competency indicators: checks understanding against predictions, oral rereads, uses context, “holds” to read further, and asks for help.

In a discussion of the Michigan and Illinois assessments, Linn (1991) concluded that the metacognitive sections “break new ground” but cautioned, “Until a good deal more research has been completed
that leads to a better understanding of the properties of these mea
sures and their construct validity, however, they are best viewed as
promising experimental approaches” (p. 193). State education offici
als would do well to heed his advice.

Additional Recommendations for New Educational Assessments

As already emphasized, the emerging consensus is that new educa
tional assessments should capture the cognitive and metacognitive processes involved in academic activities such as
emphasized the importance of focusing on metacognitive processes,
but asserted that there is a measurement problem because process is
not usually available to direct measurement. Her own instrument, the
LASSI, is an indirect form of assessment in that it relies on self-
reports, as do most of the psychometrically validated tools. We would
disagree with Weinstein’s pessimism on the feasibility of measuring
process directly, however, as would many others. For example, Tay-
lor (1987), Linn (1991), and Nason and Zabrucky (1988) advocated the
use of the computer for assessing cognitive and metacognitive pro-
cesses. The computer can continuously monitor and record all re-
sponses, adapt to the student’s responses, and make accurate time
measurements. In addition, as Taylor noted, tasks can be designed
that require the student to externalize processing steps. For example,
a list of strategies could be displayed on a main menu; the student
selects one and the computer records which was selected and when.

It also appears that the approach Meltzer (1994) is taking to
develop process measurements is a good step in the right direction
(but see criticisms by Torgesen, 1994). She seeks to “assess the
students’ metacognitive strategies and ability to coordinate the mul-
tiple subskills and strategies necessary for effective learning” (p.
594). Her recommendation is to use tasks that assess the ability to
access, use, and monitor strategies in multiple domains, academic
and nonacademic.

Consensus is also emerging for assessments to provide opportu
nities for reflection on cognitive processing (Valencia et al., 1994).
Consider the endorsement of this view that appeared in the report of
The Presidential Task Force—Learner Centered Psychological Prin-
ciples: Guidelines for School Redesign and Reform (1993); effective
assessment should promote “students’ self-reflection on their growth
by providing opportunities for self-assessment and thoughtful feed-
back on learning progress” (p. 13).
CONCLUSIONS AND RECOMMENDATIONS FOR THE FUTURE OF METACOGNITIVE ASSESSMENT

Where do we now stand with respect to the measurement of metacognition? What are our future prospects? In 1991 Paris wrote that during the past 10 years “there have been great strides made” in the assessment of metacognition in the domain of reading (p. 45). Although we are perhaps not as sanguine as he is, we agree that progress has been made. This is not so much progress in developing instruments that have been validated psychometrically, but rather in the emerging consensus that process measures rather than product measures are needed in educational assessment in general and that metacognitive assessments have their place in this new wave of testing. Throughout this chapter, we have included quotations from leading scholars in psychology and education that reflect these views. In this section we make some closing observations and recommendations regarding metacognitive assessment, addressing such issues as the value of converging evidence, domain specificity, evaluation criteria, and uses to which metacognitive assessments are put.

On the Value of Converging Evidence

That we are still far from having adequate tools for measuring metacognition is clear. One solution to the problem of measurement is to use as many methods as possible with each student. This recommendation for converging evidence is not new, having been made by Baker and Brown (1984); Garner (1988); Rushton et al. (1983); White (1988); and Weinstein and Meyer (1991), among others. However, it is sufficiently important to bear reiterating. Many investigators today do in fact use a combination of measures to obtain converging evidence. As White wrote, “Though each method is weak, the constellation of evidence from them will be more reliable and valid than each alone” (p. 74). If different measures are used that do not share the same sources of error, and the same conclusions are reached, we can be more confident that we have measured what we set out to measure. The need for obtaining converging evidence is perhaps even greater in applied settings, where the stakes to the student are higher, than it is in basic research. Recommendations to collect multiple measures occasionally appear in the literature for teachers and practitioners (e.g., Yochum & Miller, 1990), but not as often as they should.
On the Issue of Domain Specificity in Metacognitive Assessment

It is generally agreed that metacognitive knowledge and control do not occur uniformly across tasks or settings, and that the likelihood of transfer from one setting to another is quite low (Baker, 1994). Studies that have included assessments of different domains of metacognition have found low correlations among domains such as metamemory, metacommunication, metareading, and social cognition (e.g., Byrd & Gholson, 1985; Kurdek & Burt, 1981). Even within a particular metacognitive domain, there are multiple independent dimensions to the construct, as has been demonstrated in metamemory research (Belmont & Borkowski, 1988; Hertzog, Hultsch, & Dixon, 1989). Given the lack of evidence of a general metacognitive ability, it is clear that assessment instruments must be tailored to the domain or domains of interest, whether for use in research or practice.

It has been observed that metacognition is often equated with higher level thinking in the educational literature (Linn, 1991; Paris, 1991), with the unfortunate consequence that metacognition might be regarded as domain general. Thus, teachers might develop curriculum units focusing on metacognition as a decontextualized skill (and indeed we have seen such a unit in a local gifted and talented sixth grade classroom) and seek assessments that are “pure” measures of metacognition. However, the consensus among researchers is that metacognitive skills should be taught in context, not as separate aspects of the curriculum. This concern has been articulated persuasively by Paris, Jacobs, and Cross (1987):

It appears that the enthusiasm surrounding metacognition has established the construct as a pinnacle of information processing. It is the most prized, most regulative, top-of-the-hierarchy component in several theories and instructional packages. This appears to us to be an erroneous aggrandizement of decontextualized knowledge. The goal of development and education is not to produce people who reflect, orchestrate, plan, revise, and evaluate their every action. (p. 238)

Metacognition is important, but it should not serve as an instructional goal in itself but rather as a means to an end (Baker, 1994; Garner, 1987; Symons, Snyder, Cariglia-Bull, & Pressley, 1989). It follows that the assessment of metacognition should also be done in context, with measures developed in conjunction with instructional programs (Jacobs & Paris, 1987).

Measurement of metacognition is made more difficult by many of the same individual difference variables that confound measure-
ment of intelligence. Torgesen (1994) has identified four: differences in information-processing capacity and basic processing efficiency; domain-specific knowledge and experience; environmental opportunities to learn appropriate executive routines, including interactions with parents and teachers; and motivational/attitudinal variables. Content-free measures of metacognition would have the potential to reduce the influence of these confounding variables. But "because executive functioning in the real world is so interdependent with knowledge structures and basic processing efficiency, one wonders if such 'decontextualized' measures of executive processes will have much value in explaining everyday performance problems or providing proper guidance for remedial efforts" (Torgesen, p. 154).

The best intervention programs are those that work within a specific context, and so, as noted above, the focus should be on the development of methods for assessing individual differences in metacognition within specific academic domains (Torgesen, 1994).

On the Criteria for Evaluating Metacognitive Assessments

In our earlier discussions of assessment instruments, we included information about validation efforts when it was available. In many cases, it has been difficult to develop instruments that met traditional criteria of reliability and validity. Linn (1991) raised the important question of whether efforts to develop psychometrically sound assessments of metacognition are in fact misguided. As he wrote, "Constructing valid assessment procedures to tap thinking processes and metacognition is certainly not an easy task, but the difficulty of the task is not the major barrier. Practical concerns about cost and efficiency, the seemingly insatiable demand to boil everything down to a single number, and the over-reliance on standard psychometric criteria to judge test reliability and validity present much more formidable barriers" (p. 204). Perhaps we should be considering other criteria in evaluating metacognitive assessment procedures, those recommended by Linn for evaluating performance assessments, such as fairness and generalizability.

Certainly there is a need for greater ecological validity in the assessment of metacognition. Paris (1991) advocated the use of authentic text and provision of full information to students about the nature and purpose of the task. Valencia and Pearson (1986) similarly suggested that metacognition might best be assessed by observing and interacting with students while they are actually engaged in "real" reading situations, pointing out limitations of group tests:
We can and should measure these skills in formats amenable to large-scale assessment. But there will always be some limitations to data gathered from group tests of metacognitive activities: (1) what students say may differ from what they do, (2) strategic readers may be too flexible and adaptive to allow us to capture their skill in a small sample of situations and options, and (3) for many readers, these strategies operate at an unconscious, automatic level inaccessible to verbalization or even reflection. In short, here is a case in which large-scale assessment may prove moderately useful for some very limited purposes and decisions; however, the assessment strategies that really count are likely to occur at the classroom or individual level. (p. 6)

On the Uses to Which Tests of Metacognition are Put

It is important to keep in mind the various purposes for tests in the assessment of metacognition as well as in other domains. Instruments that are used in basic research are designed to answer particular questions and usually have standardized procedures. They may not be practically useful, but they may lead to the development of instruments useful in practice. In contrast, tests that are designed for diagnostic purposes need to give information that can be easily translated into educational terms (Taylor, 1987). The distinction made by Meltzer (1994) between measurement and assessment is relevant: Assessment is a broader and more inclusive term in that it entails goals and objectives, including identification of the what, how, and why of learning, and prescription, including directions for intervention and instruction. We have been using the terms interchangeably in this paper, but in reality, much of the basic research on metacognition is concerned with measurement, whereas in school settings assessment is primary.

There are variations across domains in the uses to which metacognitive tests are put. For example, as we have seen, assessments of metacognition in reading have been widely used in educational settings as well as research settings. Numerous articles for teachers and for school psychologists have offered recommendations as to how and why metacognitive aspects of reading should be assessed. In contrast, assessments of metamemory are almost exclusively the province of the research community. It is rare to see articles for practitioners calling for tests of metamemory in school settings. Is this perhaps because there is less perceived need for metamemory assessments in school? Or is it that basic researchers are refraining from putting research instruments into schools until the construct of metamemory and its measurement are more fully validated?
Another relevant issue concerns whether metacognitive assessments in applied settings should be used for diagnostic, summative, or comparative purposes. Paris (1991) cautioned teachers, "Because the goal of increasing children's metacognition about reading is only an intermediate step in the development of literacy, measures of metacognition should be diagnostic rather than summative" (p. 38). And Wittrock (1991) argued that the assessments should be for diagnostic rather than comparative purposes. He called for a new type of test that would provide diagnostic information about a student's preconceptions, learning strategies, metacognition, and affective thought processes. It would not be used to provide comparative information but rather would help provide information relevant to the diagnosis of student learning and to the design and improvement of classroom teaching by increasing teachers' understanding of these processes.

Is it time for measures of metacognition to assume a place in formal diagnostic assessment batteries? Torgesen (1994) thinks not. In fact, he offered a provocative suggestion on "how to prevent the assimilation of these measures into assessment practices for children with learning disabilities: avoid providing good norms for the measures so that they remain within a research experimental context" (p. 157). He argued that the first priority is to examine the construct validity of the measures. As we have seen, however, the assimilation may already be beginning, as witnessed by the many articles written for school psychologists on metacognitive assessment. Perhaps the findings of Farrell et al. (1989) that school psychologists are not typically assessing metacognition, at least not in England, should be seen in a positive light.

Is it time for measures of metacognition to assume a place in assessment of educational progress? As Linn (1991) argued, tests signal what is important to teachers, parents, students, and policymakers, and if these constituencies are to see that teaching metacognitive skills is important, then metacognition needs to have its place in tests. Tests, like it or not, drive instruction. Usually educators decry the practice of teaching to the test; this is of course a problem if there are specific facts that the student is to master that are assessed in standardized multiple-choice formats. This is no less true if it is a question about the strategies readers should use when they are having difficulty understanding than if it is a question about the date a historical event took place. As Kirby and Moore (1987) argued, "Instruction in metacognitive awareness, without any practical skill or strategy development, would be unlikely to improve [reading] skills or to serve any other useful function" (p. 135). A student can just as easily memorize metacognitive "facts" as histori-
cal facts. But teaching to the test is much less of a problem if the processes of thinking, reasoning, and problem solving, and the metacognitive strategies that facilitate those processes, are the focus of instruction and assessment.

Summary

Many researchers have been concerned about the uncritical acceptance of the construct of metacognition and the approaches that have been used to measure it. In this chapter, we examined measurement issues from a variety of different perspectives, beginning with a consideration of definition. Some of the difficulty in developing solid measures of metacognition stems from the differences in the way metacognition has been defined by those who study it. We define metacognition as knowledge and control of cognition, and so we discuss issues relevant to assessment of both knowledge and control. Measures discussed in detail were verbal reports (interviews, questionnaires, and think-alouds) and error detection (used most frequently in studies of comprehension monitoring). Despite their limitations, verbal reports are valuable sources of information and continue to have an important place in the assessment of metacognition. The limitations of error detection approaches have been well documented; although this paradigm has been informative, we believe it is time to focus on more ecologically valid indices.

Some researchers have sought to develop standardized instruments for assessing metacognition that are theoretically motivated and that meet psychometric criteria of reliability and validity. We selectively discussed instruments assessing metamemory assessment, metacognitive knowledge about reading, learning and study strategies, and problem solving. Although we now have a handful of instruments with reliability that is adequate for research purposes, none are sufficiently solid that they should be used for formal assessment in school or clinical settings.

Many articles have been written for teachers and school psychologists suggesting ways for them to assess metacognition in their students. We discussed the prescriptive advice given to practitioners for how they might assess metacognition using interviews, think-alouds, error detection, and process measures, and we expressed our reservations about the uncritical presentation of measures with questionable reliability and validity. Researchers who attempt to translate research into practice have an ethical obligation to frame their recommendations responsibly, providing concrete information on the limitations of the measures.
New modes of intellectual and educational assessment are being developed that focus on the processes of cognitive activity rather than the products and that measure what students do in more ecologically valid contexts. There is a growing demand to make assessment practices more in line with current views of learning and instruction. We briefly discussed the place of metacognition in some of these alternative assessments, including dynamic assessments, portfolio assessments, and performance assessments. Though these approaches are promising, they are in need of additional validation.

In the final section of the chapter, we stressed the value of converging evidence in the assessment of metacognition, the evidence that metacognitive skills should be taught and therefore assessed in context, as domain-specific rather than domain-general skills, and raised questions regarding evaluation criteria for metacognitive assessments and the uses to which such assessments are put. We do not yet have solid answers to these important questions, but we hope that the issues addressed in this paper, along with the contributions of the other participants in this timely symposium, will serve to stimulate further dialogue among researchers, educators, and policymakers about the future of metacognitive assessment.

REFERENCES


3. METACOGNITION IN CHILDREN AND ADULTS


assessments. In S. W. Valencia, E. H. Hiebert, & P. P. Afflerbach (Eds.), 
Authentic reading assessment: Practices and possibilities (pp. 255-276). 
Newark, DE: International Reading Association.

Kirby, J. R., & Moore, P. J. (1987). Metacognitive awareness about 
reading and its relation to reading ability. *Journal of Psychoeducational 
Assessment, 2*, 119-137.

study of children's knowledge about memory. *Monographs of the 
Society for Research in Child Development, 40*(1).

children's metacognitive skills: Generality and cognitive correlates. 
*Merrill-Palmer Quarterly, 27*, 287-305.

Kurtz, B. E., Reid, M. K., Borkowski, J. G., & Cavanaugh, J. C. 
*Bulletin of the Psychonomic Society, 19*, 137-140.


New York: Addison-Wesley.

York: The Guilford Press.

testing. In L. Idol & B. F. Jones (Eds.), *Educational values and cognitive 
instruction: Implications for reform* (pp. 179-208). Hillsdale, NJ: Lawrence 
Erlbaum Associates.

of task-related learning behaviors: Attention to task and metacognition. 


a scale used to measure metacognitive reading awareness. *Journal of 


Feagens, E. J. Short, & L. J. Meltzer (Eds.), *Subtypes of learning*


Assessing Metacognitive Knowledge Monitoring

Sigmund Tobias
Fordham University

Howard Everson
The College Board

Metacognition has been defined as the ability to monitor, evaluate, and make plans for one's learning (Flavell, 1979; Brown, 1980). Research has shown that learners with effective metacognitive skills are more capable of making accurate estimates of what they know and do not know, of monitoring and evaluating their on-going learning activities, and of developing plans and selecting strategies for learning new material. A large body of literature, reviewed in the other chapters of this volume, has reported differences in metacognitive abilities between learning disabled and regular students, as well as between generally capable learners and their less able counterparts. This research clearly indicates that metacognitive abilities are critically important for effective learning.

Metacognitive processes are usually divided (Pintrich, Wolters, & Baxter, this volume) into three components: knowledge about metacognition, monitoring of metacognitive processes, and control of those processes. The research described in this chapter concentrates

1Preparation of this chapter was supported by The College Board, as were Studies II–VII. Parts of this chapter were prepared while the first author was a Visiting Faculty Fellow at the Navy Personnel Research and Development Center in San Diego, in a program sponsored by the American Association for Engineering Education and the U.S. Navy.
on the monitoring component of metacognition, specifically students’ abilities to monitor their learning by differentiating between the known and unknown. It is assumed that effective control of learning cannot occur in the absence of accurate monitoring. If students cannot distinguish between what they know and do not know, they can hardly be expected to exercise control over their learning activities, or to select appropriate strategies to attain their goals.

INTRODUCTION

Our concern with assessing knowledge and/or ability monitoring is based on reasoning that it is a crucial component of most learning and training contexts. In such situations the learner usually has to master a great deal of new knowledge. Therefore, those who accurately distinguish between what they have already learned and what is yet to be acquired have an important advantage, because they can refrain from studying material that has already been mastered, or merely review it briefly. Such students can then devote most of their time and energies to new, unfamiliar materials. In contrast, those with less effective knowledge monitoring processes are likely to allocate their time and resources less effectively and spend valuable time studying what is known at the expense of the unfamiliar material and, consequently, have greater difficulty mastering new subjects. For these reasons, the program of research described in this chapter concentrated on the development of a procedure to assess students’ abilities to monitor their knowledge, and differentiate between what they believe they know and do not know and what they actually know and do not know.

The purposes of this chapter are to describe the metacognitive knowledge monitoring assessment (KMA) we have developed, and to report on a program of research—12 studies in all—that relate scores on the procedure to reading comprehension, problem solving in mathematics, and, more generally, to learning in school settings. Other studies also related scores on the KMA to such variables as anxiety, interest, and need for feedback, and examined the usefulness of the procedure in differentiating among learning disabled, attention deficit hyperactive, and students without special educational needs. All of the studies reported in this chapter used the KMA, a procedure that may be administered using paper and pencil or via computer. The procedure can also be scored objectively and, unlike other assessments of metacognitive processes, it does not rely solely on self-reports of cognitive processing.
Assessing Metacognition

Despite its importance in meaningful human learning, the assessment of metacognition has proven to be both difficult and time-consuming (Pintrich et al., this volume). Metacognition, as a higher order executive process (Borkowski, this volume; 1995), monitors and coordinates the cognitive processes employed during learning. As can be expected, there are considerable difficulties in assessing such higher level processes. Metacognition is usually assessed in two principal ways: observations of students’ performance or by self-report inventories. Some problems associated with each of these forms of assessment are described below.

Observation and Verbal Report

Assessing metacognition by observation and verbal reports usually requires all of the following: (a) that students work on some task individually; (b) that their performance is carefully observed; and (c) that their performance is recorded in some way (notes taken by observers or audio/videotapes). Often a number of additional steps are required before a rating of metacognition can be made, including detailed interviews of students, the development of “think aloud” protocols collected as students work on a learning task, and the recording of students’ introspective reports. Multiple raters are usually needed to inspect both the records of the performance and the interviews, or introspection protocols, before a sound rating of metacognition can be made (Meichenbaum, Burland, Gruson, & Cameron, 1985).

Referring to this approach, Royer, Cisero, and Carlo (1993) noted that: “The process of collecting, scoring, and analyzing protocol data is extremely labor intensive” (p. 203). The resources for such work are rarely available in most instructional situations or in many university-based research programs. Pressley’s work (this volume; Pressley & Afflerbach, 1995) provides a good example of the complexities of conducting protocol analysis, and Baker and Cerro’s chapter (this volume) also discusses some problems with this approach, especially as it pertains to the use of error detection for assessing metacognition.

Labor intensive practices such as those described above make it difficult to evaluate metacognition in many instructionally relevant settings, including secondary and post-secondary schools, as well as training environments in business-industry, governmental agencies, or in the military. In view of these difficulties it is not surprising that most metacognitive research is usually conducted in elementary and some secondary school settings where the time of those participating
in the research can easily be diverted for the research effort. Of course, substantial resources still have to be devoted to enable researchers to collect such metacognitive data.

Self-Report

A number of self-report measures of metacognition (Everson, Hartman, Tobias, & Gourgey, 1991; Jacobs & Paris, 1987; O'Neil, 1991; Pintrich, Smith, Garcia, & McKeachie, 1991; Schraw & Dennison, 1994) have been developed and are widely used. Such questionnaires have the advantage of being easily administered to groups and may be scored rapidly and objectively. Self-report scales usually ask respondents to select from a set of printed choices the cognitive processes and strategies they use while learning from instruction. Such scales put a premium on effective reading abilities and, therefore, are not usually suitable for use with younger or early elementary school children.

Unfortunately, the use of self-report measures in assessing a complex process such as metacognition raises a variety of questions, including some of the following: Because metacognition involves the monitoring, evaluation, and coordination of cognitive processes, are students aware of the processes used during learning? Further, are students able to describe and report on metacognitive processes used, even by merely selecting from available alternatives on a multiple-choice scale? Finally, there is the question of whether students report honestly on the processes. Although the truthfulness of students' answers is always an issue with self-reports, it may apply especially to reports of cognitive processes used during learning because students at any level are probably reluctant to admit that they may be relatively casual during their attempts to complete school assignments. Of course, these concerns are minimized if appraisals of any construct, and evaluations of metacognition in particular, do not rely on self-reports.

Rationale for Assessing Knowledge Monitoring

Each of the studies reported in this chapter employed a technique for assessing metacognitive monitoring that simultaneously evaluated students' self-reports of their declarative word knowledge, or their procedural math problem-solving ability, and their demonstrated knowledge or ability. The basic strategy is to assess knowledge monitoring by evaluating the discrepancy between students' estimates and their actual (determined by performance on a test) knowledge or ability. On the KMA, students are first asked to estimate their knowledge or ability to solve mathematical problems.
The actual knowledge or problem-solving ability is subsequently assessed by administering an objectively scored test, most frequently in multiple-choice format. The discrepancies between students' estimates and their actual knowledge are used as an index of the accuracy of students' metacognitive knowledge monitoring abilities.

The KMA generates four scores that reflect the relationship between students' knowledge estimates and their test performance. Two scores indicate that students estimated knowing an item, or being able to solve a problem, (a) and answered the question correctly on a test (abbreviated as + +), (b) or answered it incorrectly (+ -). Two further scores are generated indicating that students estimated that they do not know an item, or are unable to solve a problem, and (c) answer it correctly (- +), or (d) incorrectly (- -). Of course, the + + and - - scores are assumed to reflect accurate knowledge monitoring judgments, and the + - and - + scores reflect inaccurate judgments.

Like other types of metacognitive measures, KMA estimates also consist, in part, of self-reports. However, such reports typically are much more readily available to students than the questions usually appearing on self-report inventories dealing with their recollections of the cognitive processes engaged in during learning, and/or how frequently the processes were used. More important, the KMA also incorporates students' actual performances on a test. Because estimated and actual performance can both be scored objectively, the procedure has a clear-cut advantage over asking students to report on their cognitive processes either in the form of protocols, or by choosing from available alternative on self-report inventories.

School assessments are often used to determine whether students learned material presented in class. Therefore, it is important to evaluate students' ability to update their knowledge and make accurate metacognitive estimates of whether the new material was learned, in addition to assessing their prior learning. Consequently, several of the studies reported below also examined students' accuracy in monitoring whether they had mastered materials after being given the opportunity to do so.

The KMA was applied to the domain of students' declarative word knowledge in 10 of the 12 studies described in this chapter. This domain was selected because of its relevance to school learning. In order to demonstrate that the procedure generalizes to other academic domains, two studies dealt with students' procedural knowledge in the area of solving mathematical problems, another important domain in school learning at all levels. Finally, the research described below also examined the relationship of KMA scores and measures of
reading comprehension, school learning, anxiety, interest, and need for feedback, as well as examining whether the KMA differentiated between regular students and those diagnosed as being either learning disabled or having an attention deficit hyperactivity disorder.

Reports of the studies are organized into different categories according to the variables examined. Because a number of the investigations dealt with multiple variables, some studies appear under more than one rubric. In such instances, a detailed report of the study is given when it is first described, and the reader is directed back to that description in subsequent, briefer references to that investigation.

KNOWLEDGE MONITORING AND READING COMPREHENSION

There has been a good deal of research demonstrating that word knowledge or vocabulary is one of the major components of reading comprehension and learning more generally (Brelan, Jones, & Jenkins, 1994; Just & Carpenter, 1987). However, few investigations studied whether the accuracy of students' estimates of their word knowledge was an important predictor of the ability to learn. If students are unable to differentiate accurately between the words they know and do not know, they must find it difficult to determine whether to slow down while reading and try to figure out the meaning of a word from the context, or go to a dictionary to have it defined, or go on in the possibly mistaken or uncertain belief that they understand the word's meaning. Such uncertainty must be reflected in reduced reading comprehension for students with inaccurate knowledge monitoring. On the other hand, being able to distinguish accurately between words students can define correctly and those they cannot should enhance their reading comprehension and their effectiveness in learning new material. Because a great deal of research on metacognition has dealt with reading comprehension, the criterion for assessing the validity of the KMA in the first two studies was to determine its relationship to measures of reading comprehension.

Study I: Estimates of Word Knowledge and Reading Comprehension

In view of the demonstrated relationships between metacognition and reading comprehension, it seemed important to evaluate the accuracy of students' monitoring of their word knowledge in a

---

2Study I was presented at the annual convention of the American Psychological Association, in San Francisco, August 1991. That paper was co-authored by S. Tobias, H. Hartman, H. Everson, and A. Gourgey. See references.
reading context. Such a setting was expected to increase the relevance of the assessment to school learning. It was also anticipated that the ability to learn new vocabulary would be an important skill for reading specifically, and school learning more generally. Furthermore, students' abilities to make accurate metacognitive assessments of whether they had actually learned the meanings of new words, given an opportunity to do so, would also seem to be an important indicator of reading comprehension. Therefore, the ability to update one's knowledge and to make metacognitive estimates of the updated knowledge were also assessed in this study.

Participants and Procedures

Participants were randomly assigned to one of two conditions. One group was asked to read a 750-word text passage, and then complete a word list and vocabulary test composed of words that had been defined explicitly or implicitly in the text. The second group received the Sentence Verification Test (SVT; Royer, Lynch, Hambleton, & Bulgarelli, 1984) rather than the text passage as a control. The text passage described the incidence and prevalence of heart disease, the risk factors for developing heart ailments, the technical terms for varying degrees of the illness, the characteristics differentiating the different degrees, and a number of ways by which the risks of developing heart disease could be reduced. It was known from prior research (Tobias, 1989; 1969) that there was a good deal of variability in participants' prior knowledge of this material.

On the word list, participants were asked to indicate, by checking off one of two blanks, whether they knew, or did not know each of 33 words. All of the words were defined, either explicitly or implicitly, in the passage previously administered to the group who read the heart disease text. When the word list was completed, students received a four-choice vocabulary test containing all of the 33 items on the word list with instructions to select the correct synonyms or definitions of the words. A number of other research instruments were also administered, as was the Descriptive Test of Language Skills, Reading, and Comprehension (DTLS; College Board, 1979), a standardized test of reading comprehension.

The text passage, word list, and vocabulary test were examined by four raters who judged whether the words were defined implicitly or explicitly in the text. The passage was revised until consensus was reached among the judges. Of the 33 words, the ratings indicated that 25 were defined implicitly (e.g., "Epidemiologists who have com-
pared the prevalence of heart disease in the United States and in other countries..." and eight words were defined explicitly (e.g., "Coronary or heart disease...").

A total of 167 freshmen at a large urban university participated in this study. The students attended a summer session program designed to familiarize them with the university and the skills needed to succeed in their studies. The group receiving the SVT consisted of 87 students, and 82 subjects read the text passage.

**Results and Discussion**

The accuracy of students' metacognitive word knowledge judgments was determined by comparing students' estimates of their knowledge with their performance on the vocabulary test. The four scores described earlier were generated: Terms checked as being known on the word list which were scored (a) correct [abbreviated as + +], or (b) wrong [shown as + - ] on the vocabulary test. Two further scores described terms students checked as being unknown on the word list and answered (c) correctly [abbreviated as - +], or (d) incorrectly [- -] on the vocabulary test.

The four KMA scores were computed for the total set of words, and also for those that were defined explicitly or implicitly. The correlations between these data and the reading comprehension subtest of the DTLS are shown in Table 1, for all subjects combined, as well as for the group receiving the Sentence Verification Procedure (SVT) and those reading the heart disease text.

The correlations in Table 1 indicate that, as expected, accurate metacognitive judgments about the number of words students thought they knew and answered correctly on the test (T+ +) had a substantial positive relationship with reading comprehension. Estimates of the number of words thought to be unknown and answered incorrectly (T− −) were negatively related to comprehension. Furthermore, and also anticipated, accurate estimates of words defined explicitly (E + + and E − −) and implicitly (I + + and I − −) were also significantly correlated (see Table 1) with comprehension, whereas the incorrect judgments (E + −, E − +, I + −, and I − +) were not. The magnitude of many of the correlation coefficients is especially impressive because the participants were relatively homogeneous with respect to ability, because they were considered to be at risk of doing poorly in school and, therefore, advised to participate in the orientation and pre-freshmen skills program they were attending.

The relationships between the KMA scores and reading comprehension were dramatically lower for students who did not read the
Table 1. Zero-Order Correlations for Selected Variables with the DTLS Reading Comprehension Score.

<table>
<thead>
<tr>
<th></th>
<th>Entire Sample</th>
<th>SVT Group</th>
<th>Heart Disease Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>T++</td>
<td>.4655**</td>
<td>.2913*</td>
<td>.6474**</td>
</tr>
<tr>
<td>T- -</td>
<td>-.4330**</td>
<td>-.3721**</td>
<td>-.5442</td>
</tr>
<tr>
<td>T+-</td>
<td>-.1803</td>
<td>-.0885</td>
<td>-.2600</td>
</tr>
<tr>
<td>T+ -</td>
<td>.0678</td>
<td>0.2027</td>
<td>0.0825</td>
</tr>
<tr>
<td>E++</td>
<td>.3263**</td>
<td>0.0808</td>
<td>0.5221**</td>
</tr>
<tr>
<td>I++</td>
<td>.4662**</td>
<td>.3185*</td>
<td>.6302**</td>
</tr>
<tr>
<td>E- -</td>
<td>-.3349**</td>
<td>-.2894*</td>
<td>-.4196</td>
</tr>
<tr>
<td>I- -</td>
<td>-.4413**</td>
<td>-.3822**</td>
<td>-.5438**</td>
</tr>
<tr>
<td>E- +</td>
<td>-.1390</td>
<td>-.1715</td>
<td>.1151</td>
</tr>
<tr>
<td>I- +</td>
<td>-.1626</td>
<td>-.0523</td>
<td>-.2827*</td>
</tr>
<tr>
<td>E+ -</td>
<td>.1586</td>
<td>.3295*</td>
<td>.0389</td>
</tr>
<tr>
<td>I+ -</td>
<td>.0140</td>
<td>.1095</td>
<td>-.0877</td>
</tr>
</tbody>
</table>

Legend: T = total score on word list task; E = words defined explicitly; I = words defined implicitly; ++ = words Ss claimed to know and got right on a vocabulary test; - - = words Ss claimed they did not know and got wrong on a vocabulary test; + - = words Ss claimed to know and got wrong on a vocabulary test; - + = words Ss claimed they did not know but got right on a vocabulary test.

As expected, students who could update their knowledge would make more accurate metacognitive judgments than the others. Operationally then, it was expected that group membership (i.e., reading the heart disease passage or the SVT) and accuracy of metacognitive judgments would have an interactive effect on reading comprehension. This hypothesis was tested by multiple regression analysis in which a binary vector for group membership (those reading the heart disease passage or the SVT), KMA score, and their product (representing the interaction term), were entered as independent variables and the reading comprehension test score was the dependent variable. The results of that analysis are shown in Table 2.
Table 2. Beta Weights and Associated $t$ Tests for all Effects on all Derived Scores.

<table>
<thead>
<tr>
<th>Score</th>
<th>Group</th>
<th>Score Beta</th>
<th>$t$</th>
<th>Group X Score Beta</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T++</td>
<td>-.10</td>
<td>-.10</td>
<td>.44</td>
<td>-.62</td>
<td>.33**</td>
</tr>
<tr>
<td>T-</td>
<td>.12</td>
<td>.12</td>
<td>.97</td>
<td>.04</td>
<td>.19</td>
</tr>
<tr>
<td>E++</td>
<td>-.42</td>
<td>-.42</td>
<td>3.00**</td>
<td>-.34</td>
<td>1.22</td>
</tr>
<tr>
<td>E-</td>
<td>.07</td>
<td>.07</td>
<td>.56</td>
<td>.03</td>
<td>.14</td>
</tr>
<tr>
<td>I++</td>
<td>-.59</td>
<td>-.59</td>
<td>3.14**</td>
<td>-.07</td>
<td>.29</td>
</tr>
<tr>
<td>I-</td>
<td>.10</td>
<td>.10</td>
<td>.84</td>
<td>.00</td>
<td>.02</td>
</tr>
</tbody>
</table>

*p<.05
**p<.01

T = Results for total word list.
E = Results for words defined explicitly.
I = Results for words defined implicitly.
++ = Words students claimed to know and got right on vocabulary test.
+ = Words students claimed to know and got wrong on vocabulary test.
- = Words students claimed not to know and got wrong on vocabulary test.
+ = Words students claimed not to know and got right on vocabulary test.

Results indicate that students who could update their word knowledge by reading the passage made significantly more accurate metacognitive judgments than those who did not have that chance. This finding is not surprising because the major skill assessed for the group reading the passage was probably the ability to infer the meaning of words, surely an important component of reading comprehension. Clearly then, the opportunity to renew word knowledge and then estimate mastery of the updated knowledge improved the relationships with reading comprehension.

Estimates and Number Correct. The metacognitive scores described above were a function of two factors: Knowledge as reflected in the number of items students answered correctly on the vocabulary test, and knowledge estimates seen from how accurately students estimated that number. One question that arises is whether students’ knowledge estimates contributed variance above and beyond their actual knowledge reflected by the total number correct, or raw score. Of course, a great deal of research has demonstrated that students’ vocabulary scores are highly related to reading comprehension and
school learning more generally (Breland, Jones, & Jenkins, 1994; Just & Carpenter, 1987). In the KMA, the raw score may be obtained by adding the ++ and -+ scores. For the monitoring procedure to be useful, it should account for more variance than the number of items students answered correctly, irrespective of their knowledge estimates. That question is examined below in the first study and for all of the other investigations described in this chapter.

The correlation between the raw score on the vocabulary test (total number of words correct) and the DTLS was .45. As Table 1 indicates, the highest relationship among the metacognitive estimates and reading ability, $r = .65$, was between the total number of words estimated to be known and actually known (T++). The difference in the magnitude of these correlations indicates that accurate estimates of students' word knowledge contributed variance above and beyond the total vocabulary score. When T++ was forced into a regression equation, the total number of words correct, irrespective of prior estimates, did not contribute enough independent variance to enter the equation. That finding confirms the differences between the two correlation coefficients described above, and indicates that the accuracy of students' estimates of their updated vocabulary knowledge were more highly related to comprehension than the actual knowledge.

The results of this first study were encouraging with respect to the construct validity of the KMA. As expected, metacognitive assessments of students' word knowledge were more substantially related to reading comprehension than the number of correct answers alone.

**Study II. Declarative Word KMA and Reading Comprehension**

The preceding study found strong relationships between metacognitive monitoring and reading comprehension in general. The purpose of the second study was to determine the KMA's relationship both to prior reading ability and some of the components of reading comprehension, such as identifying words in context, understanding meaning, and understanding the writer's tone and assumptions. The four basic KMA scores appeared to have some similarity to the phenomena studied in signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 1991), which separates phenomena into signal and noise components. Therefore, a further purpose of this study was to examine whether the signal

---

3This study, by Howard Everson, Ivan Smolilaka, and Sigmund Tobias was published in *Stress, Anxiety, and Coping*, 1994. See References.
detection paradigm could define more useful scores than the ones used in Study I. Finally, relationships between KMA scores and measures of test anxiety were examined, and are reported later in this chapter.

Participants and Procedures

The word list and vocabulary test used in the first study were administered to students, together with the Worry subscale of the Test Anxiety Inventory (Spielberger et al., 1980), and the Descriptive Test of Language Skills, Reading, and Comprehension (College Board, 1989) that contained three subscales: identifying words in context, understanding meaning, and understanding the writer's tone and assumptions. An archival index of reading ability was obtained from the participants' school records. The participants were 117 undergraduates from a large urban university, 65% were women.

Results and Discussion

Knowledge monitoring ability was assessed by computing "hits" (i.e., the number of words each participant claimed to know and subsequently identified correctly on the vocabulary task or conversely said they did not know and failed to identify correctly on the subsequent vocabulary task), and "false alarms" (i.e., the number of words each claimed to know but did not correctly identify, and those claimed to be unknown yet correctly identified). Using signal detection theory, these two indices were transformed into a d' index that provides an estimate of metacognitive sensitivity and B, an index that provides an estimate of the participants' response bias. The reliability estimate (Cronbach, 1951) of these two indices was .78.

In general, the more capable readers demonstrated higher levels of metacognitive ability. The correlations of knowledge monitoring ability—as measured by the d' index—with prior reading ability and the experimental measure of reading comprehension were .35 and .39, respectively. Moreover, hierarchical multiple regression analyses permitted us to isolate the effects of metacognitive ability on reading test performance, once prior reading ability and anxious worry were controlled statistically. These analyses suggested that metacognitive ability was positively related to reading test performance (B = .17, t = -2.23, p = .03). Similarly, the correlations with the reading test's subscales measuring vocabulary in context, literal interpretation of text, and understanding the writers' tone and assumptions were .32, .43, and .26 respectively.
Contrasts with Study I

In Study II the text in which all of the vocabulary words were defined was not administered. The correlation of .35 between the d’ score and reading comprehension was similar to the correlation of .29 (see Table 1) found in the first study between T++ and reading comprehension for those students who did not read the text passage. Of course that relationship is much lower than the correlation of .65 found in Study I between the same variables for students reading the passage. Clearly then, these two studies suggest that the metacognitive word knowledge scores derived from the KMA had a strong, consistent relationship with standardized measures of reading comprehension and, further, that the opportunity to renew word knowledge and re-estimate mastery of the updated knowledge improved the relationships with reading comprehension.

KNOWLEDGE MONITORING AND SCHOOL LEARNING

The first two studies were encouraging with respect to the relationship of the declarative word KMA to reading comprehension. The results of these investigations indicated that metacognitive estimates were closely related to competence in the domain in which students’ estimates of knowledge were obtained (i.e., reading). One purpose of the studies described below was to examine whether the declarative word KMA was related to a more distant domain than the one in which the assessment occurred, such as learning in school. The expected relationship with school learning seemed reasonable because accurate estimates of one’s knowledge should make it easier to acquire the large amounts of new information taught in such settings. Four studies dealing with these questions are described below. Furthermore, because the vocabulary and text passage dealt largely with familiar issues and had a minimal technical vocabulary, the task of inferring the meanings of unknown words from the passage, or estimating one’s word knowledge seemed most similar to learning in courses that rely largely on conventional vocabulary, rather than introducing a large set of new technical terms. Therefore, it seemed likely that declarative word KMA scores should be more closely related to students’ learning in English and Humanities courses than in others.

Another purpose of the succeeding studies was to extend the research on metacognitive knowledge monitoring to the learning of students in secondary and post-secondary schools. As mentioned above, much of the research dealing with metacognition has been
conducted in elementary schools, and to a lesser degree in secondary or post-secondary school settings. Two of the succeeding studies examined the relationship of the knowledge monitoring procedure to students' overall achievement in college, and to their learning in different content domains, and two others used high school students and those who dropped out of school.

Study III. Knowledge Monitoring and College Learning

Students acquire a great deal of new knowledge in secondary and post-secondary schools. Therefore, their ability to estimate whether they have mastered either previously learned content or new material seemed to be an important characteristic of effective learners, especially in college. Accurate monitoring of new learning should enable students with effective knowledge monitoring strategies to concentrate on new materials and skim over familiar content. On the other hand, students with less effective knowledge monitoring may waste time practicing or reviewing what they already know, rather than zeroing in on new material or updating partially learned content. Therefore, Studies III and IV asked students to estimate their vocabulary knowledge twice: the first time to assess their prior learning, and the second to determine their ability to update prior learning. It was assumed that students' accuracy in estimating their word knowledge after having the chance to update it would be more closely related to college learning reflected in their grade point average (GPA) than to estimates of prior learning.

The word list, vocabulary test, and text materials used in the two studies reported above contained a much larger set of explicitly defined words compared to those defined implicitly. It was reasoned that implicit definitions might be especially important for college learning, where students frequently had to infer the meanings of new words from context. Therefore, the materials were modified to increase the number of implicitly defined words.

Participants and Procedures

The word list, vocabulary test, and text passage were revised to contain an equal number of target words that were defined explicitly and implicitly in the text passage. The expository text used in one of the prior studies was revised and a narrative version of the same passage was developed in order to examine the effect of situational

---

“A paper based on Studies III and IV was presented at the annual meeting of the American Educational Research Association, San Francisco, CA, April 1995.”
interest on metacognitive knowledge monitoring (findings dealing with interest are discussed later in the chapter).

The word list and vocabulary test contained 38 words, half were explicitly defined and the others received implicit definitions. Types of definitions were determined by two independent judges who rated all words. Disagreements were resolved by revising the passage until agreement was reached. Because these materials are used in six of the studies described in this chapter, a sample, consisting of the first page of the materials, is shown in Figures 1-3.

The word list and vocabulary test (alpha reliability = .80) were administered in a first session. Students were then randomly assigned to one of the two versions of the text in a second session, followed by a re-administration of the word list and vocabulary test. The materials were administered during students’ classes in the presence of their instructors.

Figure 1. Word List for Knowledge Monitoring Procedure.

Please indicate whether you know, or do not know each of the words listed below, by checking the appropriate space.

<table>
<thead>
<tr>
<th>Word</th>
<th>Know</th>
<th>Do not know</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abuse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acute</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ascribed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attenuate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attributed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benign</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cholesterol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coronary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deterrent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficacy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emanating</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epidemiology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Esoteric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Etiology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fatalities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gravity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guarded</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implicated</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 2. Example of Vocabulary Items for the Knowledge Monitoring Procedure

For each word check the space which means most nearly the same thing as the first word.

1) Prevalent  
   _a) stronger _a) refer _a) tooth decay
   _b) winning _b) written _b) particle
   _c) frequent _c) question _c) rule violation
   _d) prior _d) bed _d) muscle death

2) Attributed  
   _a) caused _a) move _a) fatty tissue
   _b) ovation _b) temporary _b) deaths
   _c) stream _c) carry _c) fateful
   _d) tax _d) train _d) take in stride

3) Optimal  
   _a) best _a) stripe _a) new cases
   _b) opening _b) divider _b) an example
   _c) eyeball _c) middlemost _c) exciting
   _d) cheerful _d) negotiate _d) event

4) Obesity  
   _a) listen _a) joke _a) listen
   _b) fat _b) eat _b) reduce
   _c) apology _c) enter _c) pay attention
   _d) obsolete _d) exit _d) try

5) Acute  
   _a) pretty _a) lasting _a) uncertain
   _b) serious _b) live _b) optimistic
   _c) heavy _c) income _c) degrees
   _d) often _d) clever _d) watchful

Please turn to next page to continue

The sample consisted of 139 students attending a large urban university, though only 84 subjects completed all the materials during two sessions. Part of the sample consisted of students entering the nursing program (N = 47, N = 33 with complete data) who were taking an orientation course in that department. The rest of the sample consisted of freshmen (N = 92, N = 51 with complete data) taking a freshman orientation course.

Results and Discussion

The correlation between total score on both administrations of the vocabulary test, based on 84 students who completed the test on both administrations, was .75. This is not a test-retest reliability coefficient
Coronary or heart disease is a major health problem among all ethnic, racial and occupational groups in the United States. In addition to coronary disease, health workers are worried about many other maladies affecting Americans, such as cancer, AIDS, and other equally serious conditions. However, compared to all other serious illnesses, coronary problems cause more than half of the total number of fatalities or deaths in the United States. To be exact, 55% of the deaths among all groups in this country, or more fatalities than for all the other illnesses combined, may be ascribed to coronary disease. Not only is coronary disease responsible for the greatest number of fatalities in this country but it is also the most prevalent, or frequent, of all the serious illnesses. That is, coronary disease is more prevalent than all the other serious conditions combined.

The incidence, that is the number of new cases, of coronary disorders is higher among men than among women for the country as a whole. The incidence of heart disorders is also higher for cigarette smokers than it is among non-smokers. A higher incidence of coronary disease among Americans is also attributed to alcoholism, drug addiction, and tobacco. The etiology, or causes, of coronary disease among Americans are not completely clear, but excessive use, or abuse of alcohol and the other substances mentioned above is often linked to coronary disease. In addition, tension, air pollution, weighing too much, and engaging in too little exercise are also implicated as causes of heart disease among people living in the United States.

The gravity of heart disease for people in general is a function of the magnitude of coronary damage. The heart is basically a muscle similar to all the others in the human body. The amount of damage to the heart muscle, or myocardium, determines the seriousness of the illness. The most serious type of damage, which is called myocardial infarction, occurs when the heart muscle dies. One major difference between the myocardium and other muscles in the human body . . .
because students read the text passage, from which the meaning of the words could be inferred, immediately before the second administration of the vocabulary test.

Students' estimated word knowledge and performance on the vocabulary test were determined for both administrations. Two scores were computed for each administration: the total number of correct [words in the + + and - - categories] and incorrect [+ - and - + categories] estimates. Preliminary analysis found no differences between students assigned to the expository or narrative text versions, or between explicitly and implicitly defined words, therefore the data for both text versions and both types of words were pooled. The correlations between the correct and incorrect estimates on both administrations of the words and students' GPA in English, Humanities, Sciences, Social Sciences, and combined GPA were computed and are shown in Table 3. Because 92 participants were freshmen in their first term of college the overall GPA for this group was based on an average of only 12.1 credits ($SD = 5.6$), whereas the nursing students had a mean of 56.4 credits ($SD = 28.3$). Therefore, the correlations are presented for each group separately, as well as for the total sample. Table 3 also shows the correlations for metacognitive estimates and raw score, number correct on the vocabulary, separately. Finally, the different number of cases in the various cells of Table 3 should also be noted.

The correlations shown in Table 3 are generally positive and frequently significant, even though they ranged from low to moderate in magnitude. The results support the concurrent validity of the procedure with respect to its relationship to learning in college. As expected, correlations between knowledge monitoring scores and GPA in English were generally highest; presumably accurately estimating word knowledge is more important in English than in other subjects. Relationships with Humanities courses and with the combined GPA were generally significant and lower than those with English grades; correlations with social science and science GPAs were generally lower, and usually not significant. The largely nonsignificant relationships with social and behavioral science courses were surprising because it had been assumed that these courses usually contained less technical or unfamiliar material and vocabulary than the sciences. Perhaps grades in these courses, like those in science, reflected greater domain specific knowledge than found in English and Humanities classes.

The significance levels of the correlations reported in Table 3 varies widely, probably as a function of at least three factors. First, the number of cases in each cell differs due to students' absence from
Table 3. Correlations Between Knowledge Monitoring Procedure Scores, Raw Scores, and Overall Grade Point Averages in Different Subject Areas.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Administration 1</th>
<th>Administration 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Raw Scr</td>
</tr>
<tr>
<td></td>
<td>Estimate</td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total GPA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n 101</td>
<td>r .20*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r .01</td>
</tr>
<tr>
<td>Freshmen</td>
<td>n 65</td>
<td>r .09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r -.25</td>
</tr>
<tr>
<td>Nurses</td>
<td>n 36</td>
<td>r .28*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r -.37*</td>
</tr>
<tr>
<td>English GPA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n 72</td>
<td>r .30**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r .19</td>
</tr>
<tr>
<td>Freshmen</td>
<td>n 53</td>
<td>r .31**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r .10</td>
</tr>
<tr>
<td>Nurses</td>
<td>n 19</td>
<td>r .25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r .33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humanities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n 82</td>
<td>r .26**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r .04</td>
</tr>
<tr>
<td>Freshmen</td>
<td>n 52</td>
<td>r .12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r -.21</td>
</tr>
<tr>
<td>Nurses</td>
<td>n 30</td>
<td>r .48**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r .40*</td>
</tr>
<tr>
<td>Science GPA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n 65</td>
<td>r .18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r -.01</td>
</tr>
<tr>
<td>Freshmen</td>
<td>n 28</td>
<td>r .11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r -.30</td>
</tr>
<tr>
<td>Nurses</td>
<td>n 37</td>
<td>r .26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r -.42*</td>
</tr>
<tr>
<td>Social Science GPA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n 64</td>
<td>r .18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r .26</td>
</tr>
<tr>
<td>Freshmen</td>
<td>n 26</td>
<td>r .15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r .10</td>
</tr>
<tr>
<td>Nurses</td>
<td>n 38</td>
<td>r .09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>r -.31</td>
</tr>
</tbody>
</table>

* = p < .05  
** = p < .01

either the first or second administration of the materials. Second, it is well known that college grades are often unreliable (Werts, Linn, & Joreskog, 1978; Willingham, Lewis, Morgan, & Ramist, 1990), reducing the magnitude of any correlations with them. Third, students completed a varying number of courses in each area, thus GPAs may have been based on one or a few courses in some fields, reducing the stability of the criterion. The reliability of the grades may have been reduced further by three factors: (a) students took dissimilar courses in each of the areas shown in Table 3; (b) when similar courses were
taken they were taught by different instructors; and (c) differences in students' major fields of study.

As expected, the correlations between knowledge monitoring scores and grades in English were generally higher, and more frequently significant, than those of any other subject. For the 84 students with complete data for both administrations of the vocabulary test, the mean total score increased from 23.3 (SD = 6.0) for the first vocabulary test to 26.0 (SD = 6.6) for the second (t(83) = 5.53, p < .001). Thus students clearly learned the meanings of some of the words after having the chance to update their word knowledge by reading the passage. However, in contrast to the results of the first study, the relationships between the metacognitive scores and grades shown in Table 3 were generally higher before students read the text passage than afterwards. The Study I findings of higher relationships with DTLS scores on the second administration of the procedure may be attributable to the use of reading comprehension scores rather than grades as criteria. Apparently, inferring the meaning of words is a more important component of reading comprehension than of school learning more generally.

It was assumed in this study that having the chance to update one's word knowledge before estimating it would be more similar to students' learning in their classes than merely estimating prior word knowledge. Therefore, relationships with grades were expected to be higher for the second administration than the first. The findings did not support these expectations. Although the increase in vocabulary score after reading the text was statistically significant, it indicated that, on average, less than three new words were learned from the text passage. Perhaps such modest acquisition was dissimilar to the amount of learning in college courses leading to lower relationships with metacognitive monitoring scores on the second administration of the procedure. Similarities between the knowledge monitoring task and school learning might have increased if students were instructed to study the passage more intensely, or asked to pay special attention while reading words they had previously seen on the vocabulary test. Such instructions may have increased the correlations with GPA for the second administration. It remains for further research to explore that possibility.

Table 3 also indicated that the correlations with number correct on the vocabulary test were generally similar to the relationships with correct knowledge monitoring estimates. Due to the varying Ns in the different cells, the significance of differences in correlations was examined with a t test developed by Hotelling (1931). For the
correlations with GPA based on both administrations, using the total group, the knowledge monitoring scores were higher seven times (one difference was significant at $p < .05$), and the correlations based on raw scores were higher three times (none significantly so). For freshmen, the correlations with knowledge monitoring procedure scores were higher twice, but not significantly so, and correlations with raw score were higher eight times (two significant at $p < .05$). Finally, for nursing students, correlations based on knowledge monitoring scores were higher five times (none significant), and relationships based on raw scores were higher five times (one significant $p < .05$). Thus, the knowledge monitoring scores appeared to add little independent variance to the relationship with grades beyond that accounted for by the number correct on the vocabulary test.

The findings for this study, in contrast to the findings of the first two investigations, suggest that the knowledge estimating procedure seems to account for little independent variance in GPA above that attributable to number correct on the vocabulary test. Conceivably the findings of low reliability for college grades (Werts, Linn, & Joreskog, 1978; Willingham, Lewis, Morgan, & Ramist, 1990), referred to above, may have contributed to these findings. The criterion in the first two studies consisted of test scores, which are much more reliable than grades.

**Study IV: Predicting College Learning from KMA Scores**

The preceding study dealt with the concurrent validity of the KMA by examining the correlations of knowledge monitoring procedure scores with students' prior learning in college. The fourth study investigated the KMA scores' predictive validity by examining whether metacognitive estimates predicted entering students' performance during their first year of college.

**Participants and Procedures**

The materials used were identical to those described in Study III. They were administered while students attended a prefreshman skills program prior to beginning their first semester of college. Learning was determined by obtaining students' GPAs at the end of their first college year in the same areas examined in the prior study: English, Humanities, Science, and Social and Behavioral Science, as well as the combined GPA. The sample consisted of 115 students (59 female) participating in a prefreshman skills program intended for students considered at risk of doing poorly in their first year of college.
Results and Discussion

The number of correct metacognitive estimates of students' word knowledge were determined. As in the prior studies, correct estimates were defined by combining the + + and - - categories. Preliminary analysis again indicated that there were no differences between the expository and narrative passages, nor between the words defined explicitly or implicitly. Therefore, these data were pooled for the succeeding analyses.

Correlation analysis was the optimal data analytic mode in the preceding study because of the large amount of missing data due to student's absences, and the varying courses in different areas taken by the freshmen and prenursing students. However, by examining whether changes in knowledge monitoring scores were accompanied by similar changes in GPA, correlations were likely to maximize errors attributable to the low reliability of grades because small changes that could alter the correlations might be attributable to error. The participants in Study IV were incoming freshmen who completed all the materials and took similar types of courses. Therefore, high and low achievement groups were created by splitting students at the GPA median in the different academic areas, and on the combined GPA, and then examining knowledge monitoring differences between the groups. Mixed between and within subjects analyses of variance were then computed to determine the significance of differences between the first and second administrations, and of differences in metacognitive estimates between groups above and below the GPA median.

A search of the college records found that 95 of the 115 students examined a year earlier had completed some courses at the school. ANOVA indicated that, as expected, students above the median GPA \((N = 48)\) made significantly more accurate overall metacognitive judgments \((\text{Mean} = 49.2, F(1, 93) = 6.42, p < .05)\) on both administrations than those below the median \((N = 47, \text{Mean} = 45.8)\); the size of that effect, determined by \(\eta^2\) (SPSS, 1993), was .065. Also as expected, there was a significant difference between the first \((\text{Mean} = 22.9)\) and second administration \((\text{Mean} = 24.5)\) of the word list and vocabulary test \((F(1,93) = 14.95, p < .01, \eta^2 = .138)\), though there was no interaction between these variables. A similar analysis was computed using the number right on both administrations of the vocabulary test as the dependent variable. That analysis indicated that the differences between the high \((\text{Mean} = 43.2)\) and low GPA group \((\text{Mean} = 39.3)\) on the vocabulary test was not significant \((F(1, 93) =\)
2.73, \( \eta^2 = .029 \), and the differences between the first (Mean = 17.7) and second administrations (Mean = 24.5) were significant \( (F(1,93) = 198.04, p < .001, \eta^2 = .68) \); again there was no interaction.

High and low groups in English, Humanities, Science, and Social Science courses were also formed by splitting the students at the GPA median in each of these content areas and examining the significance of differences on the number of correct metacognitive estimates. In English, the overall differences in metacognitive accuracy between students above (Mean = 48.9) and below the median (Mean = 45.4) were significant \( (F(1,82) = 6.18, p < .02, \eta^2 = .07) \), as were the differences between the first (Mean = 45.6) and second administrations (Mean = 48.7, \( F(1,82) = 11.92, p < .01; \eta^2 = .127) \). Furthermore, there was an interaction between groups and administrations \( (F(1,82) = 4.41, p < .05; \eta^2 = .051) \). The interaction, shown in Figure 4, suggests that although both groups increased their accuracy from the first to the second vocabulary test in estimating which words were known and unknown, higher achieving students had greater gains than the others. A similar analysis was

![Figure 4. Interaction of GPA Groups, Hits, and Administration.](image_url)
computed for number correct on both vocabulary test administrations. The finding indicated a slightly smaller difference between the high (Mean = 42.9) and low GPA group (Mean = 38.9, $F(1,82) = 5.43; \eta^2 = .062$) than obtained when the metacognitive scores were used, but a stronger effect for differences between first (Mean = 18.0) and second administrations (Mean = 23.6, $F(1,82) = 169, p < .001; \eta^2 = .673$); there was no evidence for interaction in these results.

Similar analyses were computed for students above and below the median in Humanities courses (Art, History, Music, Philosophy, World Civilization, World Humanities, and World Arts). Differences between High (Mean = 49.4) and Low Humanities GPA groups (Mean=45.3) were also significant ($F(1, 81) = 7.96, p < .01; \eta^2 = .089$), as were the differences between first (Mean = 23.0) and second administrations (Mean = 24.5, $F(1, 81) = 9.94, p < .001; \eta^2 = .109$), there was no interaction. The same type of analysis was also computed for number correct on the first and second vocabulary test again revealing somewhat smaller differences between the high (Mean=43.1) and low groups (Mean = 39.0, $F(1, 81) = 4.18, p < .05; \eta^2 = .049$) and larger differences between the first (Mean = 17.8) and second administration (Mean = 23.4, $F(1, 81) = 179.2, p < .001; \eta^2 = .689$) than the results for knowledge monitoring scores. There were no significant differences between the Science or Social and Behavioral Science GPA groups using either the knowledge monitoring procedure or raw scores.

The relationships between metacognitive scores and GPA a year later were generally similar to those reported in Study III, supporting the predictive validity of the KMA scores. Unlike the prior study, in which both knowledge monitoring and raw scores had fairly similar patterns of relationships, the metacognitive scores had a significant effect on overall GPA, whereas the raw scores did not. Furthermore, the knowledge monitoring scores accounted for more variance than the number right in two of three other comparisons, supporting the construct validity of the procedure.

Several factors are likely to have reduced the magnitude of the effects and the generalizability of the results to other college groups. As was the case in the first study, the participants in the pre freshmen program were considered to be at risk for poor performance in college. This factor may have reduced the range of college achievement for the sample and, therefore, may also have reduced knowledge monitoring differences between the groups. Furthermore, even though data were not collected in sections of the pre freshmen skills program devoted exclusively to English as a Second Language (ESL),
some of the students were signed up for both ESL and other skills sections, and thus ended up as part of the sample. The presence of nonnative English speakers could also have reduced the variability among participants and narrowed group differences in this study. Further research limited to native English speakers, who are more heterogeneous in academic skills than the present sample, is needed to determine whether metacognitive differences between low and high achieving students are greater than those reported here.

In general, KMA scores seemed to differentiate the more capable students, whose grades were above the median, from those less able more successfully than did the raw scores, replicating the findings of Studies I and III. The knowledge monitoring scores accounted for anywhere from 1% to 4% more variance than similar analyses using the raw score. It was also interesting to note that the analysis of raw score differences between the first and second vocabulary test administrations always accounted for substantially more variance than did a similar analysis based on knowledge monitoring scores. The latter finding is reasonable and supports the construct validity of the procedure because most students learned some new words from the text passage, though their knowledge monitoring was not equally enhanced. However, it should be noted that the results for English grades indicated that there were greater increases in knowledge monitoring accuracy for capable students than for their less able peers (see Figure 4). These findings suggest that although all students increased both their knowledge and knowledge monitoring accuracy from first to second administration, the increases in metacognitive knowledge monitoring accuracy were greater for more capable students (i.e., those whose English grades were above the median). Apparently such students’ metacognitive skills improved to a greater degree than those of their less able colleagues.

It should be noted that many of the students in this sample took less than a full-time schedule of courses. That is likely to have decreased the reliability of the GPA because it was based on fewer courses and credits than is usually the case after a year of college. This factor may also limit the generalizability of the results to other groups, in addition to reducing the magnitude of the findings by decreasing the potential variability of the GPA. Therefore, in order to increase both the reliability and variability of the criterion, it would be useful to investigate the predictive validity of the knowledge monitoring procedure in settings with a greater percentage of full-time students.
Study V: Knowledge Monitoring and Learning Among Vocational High School Students\textsuperscript{5}

All of the prior studies used college students as subjects. College students are probably more academically oriented than those attending vocational high schools, and consequently more likely to be reflective about their declarative word knowledge and in turn, likely to make more accurate estimates of what they know and do not know. Therefore, one purpose of the next study was to examine the relationship of the knowledge monitoring procedure for students attending a vocational high school. This study also examined the relationship between metacognitive monitoring scores and students' estimates of their performance, as well as their test anxiety. The results dealing with those variables are discussed in later sections of this chapter.

Participants and Procedures

This study employed the word list and vocabulary test described in the two preceding studies; the text passage was not used. Students were tested in one of their regular school classes. In addition some anxiety scales were administered and students were also asked to estimate their grades on tests given in one of their vocational classes. Students' overall GPAs were obtained from the school's permanent records.

All of the participants attended a vocational high school in a large urban school system. A total of 61 students (59 male) participated in this study. The students' ages ranged from 16-19.

Results and Discussion

Students were split at the GPA median, and two multivariate analyses of variance (MANOVA) computed. The first examined differences between the high and low GPA groups on the accuracy of students' knowledge estimates (using the + +, + -, - +, and - - scores) and the second analysis examined group differences in student word knowledge (the sum of ++ and + - scores equal the number correct on the vocabulary test). MANOVA indicated that the overall knowledge monitoring differences between the high and low GPA groups were significant (Transformed Wilks lambda $F(3,57) = 3.17$, $p < .05$, effect size = .143). Univariate analyses found that only the difference between the high (Mean = 17.8) and low (Mean = 14.4) GPA groups on the + + scores were significant ($F(1,59) = 9.35$, $p < .01$).

\textsuperscript{5}The data for this study were collected by Deno Charalambous.
The MANOVA computed on group differences in the number correct on the vocabulary also indicated a significant difference between the groups (Transformed Wilks lambda $F(2,58) = 5.35, p < .01$, effect size $= .156$). Univariate analyses found that the differences in $+$ scores were the same as in the preceding analysis; however, in this analysis group differences between the high (Mean = 3.5) and low (Mean = 5.2) GPA groups on the $-$ scores were also significant ($F(2,58) = 5.59, p < .05$). As expected, the results indicated that the more capable students estimated and actually knew more words than those with lower GPAs; the latter group estimated not knowing more words than the students who were above the GPA median.

The significant differences between the two GPA groups replicate the results of the two prior studies dealing with college GPAs and confirms the relationships between metacognitive knowledge estimates and school learning. The results of the second analysis fail to support the additional importance of obtaining students’ estimates of their knowledge because the differences between the GPA groups on their actual vocabulary knowledge were also significant and slightly greater than the differences in knowledge monitoring. The word list and vocabulary test were developed for a college population; perhaps these materials were so unfamiliar to these vocational high school students that their estimates were based on little more than chance.

Study VI. Knowledge Monitoring Among High School Dropouts

The high percentage of students who drop out of high school before graduating is a major problem, especially in times when entry level employees for most positions in business and industry call for greater levels of skill than ever before. At a time when the advent of the information super-highway is beginning to redefine the job functions of lower and mid-level workers in business and industry, it is vital that students complete a secondary school education in order to have better chances of finding employment. There is little reason for optimism in that respect as Mann (1986) reported that “A national estimate suggests that 25 percent of fifth graders will not make it through high school graduation” (Mann, p. 309).

There are many reasons for students dropping out of school, but Tanner (1990) suggests that “School based reasons are the most important self-reported explanation of dropping-out for all groups of adolescents” (p. 80). Chief among the school-based reasons is poor performance in school. When asked why they had dropped out of

---

6The data for this study were collected by Heather Gerrity.
school, more than one third of the students said "Because I had bad grades," or "because I did not like school" (Mann, 1986, p. 309). These findings were substantially similar to those reported by Ekstrom, Goertz, Pollack, and Rock (1986). Therefore it seemed reasonable to assume that students who dropped out of high school would have lower metacognitive knowledge monitoring abilities than regular students. This expectation was examined in Study VI.

Participants and Procedures

The word list and vocabulary test employed in Studies II–V were administered, together with some test anxiety scales, described later in this chapter. The text passage was not used.

A total of 89 subjects participated. The dropout group consisted of 42 individuals (14 female) who had dropped out of high school and were attending a General Equivalency Diploma program. The continuing student group consisted of 47 students (16 female) who had a school GPA of at least B-. None of these students had given any indication that they were at risk of dropping out of school.

Results and Discussion

Two MANOVAs, identical to those of the preceding study, were computed to determine the significance of differences between the high school dropouts and continuing students. The first analysis found significant overall group differences (Transformed Wilks lambda $F (3,79) = 4.08, p > .01$, effect size = .134) in knowledge estimates (+ +, + -, and - - scores). Univariate analyses indicated that the dropout group (Mean = 12.7) differed from the continuing students (Mean = 16.2) on the + + scores ($F(1,81) = 8.83, p < .01$) and on the + - scores (dropout Mean = 10.6, continuing students 8.5; $F(1,81) = 6.11, p < .02$). A similar analysis of actual knowledge (+ + and - + scores) also indicated significant, though somewhat smaller group differences (Transformed Wilks lambda $F (2,80) = 4.61, p < .01$, effect size = .103). Univariate analyses indicated that only the + + score difference was significant.

The results indicate that, as expected, students who dropped out of high school had less effective knowledge monitoring abilities than did continuing students. Analysis of raw score differences yielded similar, though somewhat smaller effects. The results suggest that the poorer knowledge monitoring abilities of students who dropped out of school may have made school work more difficult for these students and contributed to poor performance, a picture that is consistent with the descriptions in the literature of school dropouts.
Summary: Knowledge Monitoring and School Learning

As expected, the four studies found significant relationships between metacognitive knowledge monitoring scores and school learning. The studies used different types of samples ranging from regular college students, those about to enter college and enrolled in a prefreshmen skills program, vocational and regular high school students, and those who dropped out of school. Because relationships with knowledge monitoring were in the expected direction for the different samples it may be inferred that the KMA has some generality across a variety of student groups. In most of the studies, the KMA scores accounted for more variance than raw vocabulary scores supporting the construct validity of the procedure.

PERFORMANCE EVALUATION AND METACOGNITIVE KNOWLEDGE MONITORING

The studies reported in the preceding section described the relationships between knowledge monitoring and school learning. The grades students receive are a function not only of their domain knowledge, but also of the standards and types of evaluations and grading procedures used by instructors. These factors potentially add error to the relationship between knowledge monitoring scores and GPA. In view of the fact that the KMA assesses students' abilities to estimate their knowledge, it was reasoned that students who were capable of accurately estimating their word knowledge on the KMA should also be more accurate in predicting their performance on examinations related to their present studies before they take them, and how well they performed on those examinations after they were completed. This section describes three studies examining these questions.

There has been some research on the prediction of performance in courses and on tests, though none of these related the predictions to metacognition or knowledge monitoring. Keefer (1971) found that college students who accurately estimated their performance achieved at a significantly higher level than less accurate estimators, and had a more positive self-concept than their low-estimating counterparts. Holen and Newhouse (1976) found that students' predictions of their grades on a course examination correlated as highly with actual performance as their GPA, and were significantly more accurate predictors than other variables, such as grades in prerequisite courses or GPA. Furthermore, students' performance predictions contributed significant
unique variance to predictions of actual final grade, above that contributed by high school and college GPA, or grades in prerequisite courses. Harris (1990) found that accurate estimators of test performance in psychology earned a significantly higher final average in introductory psychology than did low and less accurate estimators.

The research on performance estimation suggests that more capable students make more accurate predictions of their performance than their less able counterparts. Because the studies described in the preceding section found that accurate KMA scores were associated with higher GPA, the findings dealing with performance estimation support the rationale that students who make accurate metacognitive assessment of their knowledge should make more accurate predictions of test scores than would less accurate students.

Study VII. Performance Estimation and Predicting Standardized Test Scores

In some of the previous studies all students responded to the metacognitive procedure before and after reading the text passage. The results indicated that metacognitive estimates before students read the text passage were somewhat more highly related to their GPAs than those obtained after reading the passage, although the opposite findings emerged in studies relating knowledge monitoring to reading comprehension. A further purpose of this study was to vary the administration of the text passage in order to examine its contribution to students’ estimates of their test performance. Furthermore, it was considered useful to examine performance on a standardized test of known reliability to reduce possible error. Studies I and II used a standardized measure of reading comprehension as the criterion, and the results relating test performance to word KMA seemed more positive than the comparisons with the less reliable student grades. Therefore, the use of a test that had demonstrated reliability (.88) in a previous study seemed desirable.

It was expected that General Psychology students who could accurately monitor their knowledge would also be more accurate in predicting their actual and estimated scores on the Advanced Placement Test (AP) in Psychology (College Entrance Examination Board, 1988) before and after completing it, and that they would also earn higher scores on the test than their less accurate peers. Finally, as suggested by other studies of student’s estimation of their performance, it was predicted that they would expect to obtain higher grades in the course in which they were registered.
Participants and Procedures

The AP Examination in Psychology (College Entrance Examination Board, 1988) was administered to students enrolled in an introductory psychology class. Students received a description of the different areas covered by the AP test and were asked to predict how many of the 100 items they could answer correctly before the test was taken, and again after it was completed (postdiction). Half the sample \((N = 39)\) was randomly assigned to read the expository version of the text passage used in the two preceding studies before the word list, and the other half \((N = 38)\) received an unrelated task, the text selection titled “Teaching the Mentally Retarded” from the Sentence Verification Technique (Royer, Carlo, Dufresne, & Mestre, 1994), and then answered the questions on that passage. The same word list and vocabulary test used in Studies II–VI were then administered to all participants.

Students were also asked to predict their final grade in the Introductory Psychology class they were taking. On this campus the accuracy of their grade predictions could not be determined because regulations for the protection of students’ privacy made it impossible to obtain information by which they could be identified.

A total of 77 students (41 females) taking the Introductory Psychology class on one of the campuses of a large urban university volunteered to participate in the study. Students could choose from a number of projects to satisfy a requirement for research participation.

Results and Discussion

More accurate metacognitive scores were expected for the group responding to the word list and vocabulary test after reading the text compared to the other group who received the SVT, which was irrelevant to the task. Surprisingly, MANOVA based on the total number of accurate estimates [+] and [−] revealed no significant differences between the groups (see Figure 5). Examination of the basic eight scores [+,−,−+,−− for both explicitly and implicitly defined words] indicated that there appeared to be some group differences (see Figure 6), but that these canceled each other out when the data were combined into total number of correct estimates.

When MANOVA was computed on six of the basic scores (the scores for the + − category for explicitly and implicitly defined words were eliminated to avoid linear dependencies) the overall differences between the groups were significant \((F(6,70) = 3.71, p < .01)\). Univariate
F tests indicated that the students who read the passage made more accurate metacognitive estimates on explicitly defined words in the + + category ($F(1,75) = 5.97, p < .02$), and had fewer explicitly defined words in the - - category, ($F(1,75) = 4.74, p < .05$).

Predictions of performance on the AP test were tested by splitting students at the median on total number of accurate metacognitive estimates [combining the + + and - -] and computing MANOVA to examine the significance of the differences on students' pre- and postdictions of their AP scores, their actual AP score, and their expected final grade in the psychology class. There were no differences on the AP test data or on the expected final grades between groups who read either the text passage or the SVT ($F(4,69) p < 1$). Differences between high and low metacognitive groups were significant ($F(4,69) = 2.83, p < .05$; effect size = .141). Univariate tests indicated that the high knowledge monitoring group obtained higher AP scores (Mean = 43.6, $F(1,72) = 7.81 p < .01$) than the low (Mean = 36.6), and that differences in expected final grade in the course just
failed of significance \(F(1,72) = 3.40, p < .10\). There was no interaction between groups who read either the text or SVT and knowledge monitoring groups.

The data were also analyzed for the number correct on the vocabulary test by splitting the groups at the median on the number correct, and computing the significance of differences on the AP and final grade data. Those results were similar to the prior analysis using knowledge monitoring scores. That is, there were no differences between the groups who had read the text or received the SVT, and there was a significant difference between groups above and below the median on prescore \(F(4,69) = 6.47, p < .01;\) effect size = .27). Univariate analysis again indicated only one significant difference on actual AP score between groups above (Mean = 45.4) and below (Mean = 34.2) the median on vocabulary score. Again there was no interaction among the variables. Unlike the prior studies, where differences in metacognitive estimates were usually greater than those on the vocabulary raw score, the effect size for these data using the vocabulary test results was larger than for the knowledge monitoring data (.27 compared to .14).

The results indicate that students high on vocabulary score and on the ability to monitor their word knowledge also obtained higher
scores on the AP exam and expected higher final grades in the course for which they were registered. The absence of group differences on predicted AP score before taking the test was not surprising because students were unfamiliar with the test, beyond being informed about the categories of knowledge covered. They had no information about the difficulty of the items, the types of preparation expected for the test, or specifically what they would be questioned on. The absence of differences on students' postdictions was a little more surprising because participants now had a much clearer idea about what the test covered. Perhaps this brief exposure to the test was inadequate to familiarize them with the domain covered by the AP test.

Study VIII: Knowledge Monitoring and Estimations of Academic Achievement

Ideally, of course, participants' performance estimates about both predicted and actual grades should have been studied in courses for which they were registered. Unlike the AP test, students should have enough information to make more reasonable predictions about their final grades in courses, based on their experience in the class, and with the subject matter, instructor, and procedures of the course. It was the purpose of this study to examine these expectations, in addition to attempting to replicate the findings for the AP data.

Participants and Procedures

The procedures were identical to those in the previous study with two exceptions. First, the predictions students made about their final grade were compared to the actual final grade obtained in the course. Second, students took 12 quizzes in this class (the instructor used the 10 highest quiz scores in the determination of the final grade) and the grades obtained on these quizzes were available as additional dependent variables.

A total of 75 students enrolled in the Introductory Psychology class participated in this study. The students received extra credit for taking part in the research.

Results and Discussion

The first set of analyses were computed to examine the consistency between the findings of this study and the preceding one. As in the prior study, a test for significance of differences between the group who read the text and the SVT on the + +, + -, and - - scores revealed no differences between the groups. When the component
scores, based on explicit and implicitly defined words, were examined, overall differences between the groups were significant $F(6, 68) = 2.57, p < .05$). Univariate analysis indicated that the group reading the text had fewer $-$ scores for explicitly defined words ($F(1, 73) = 7.69, p < .01$) and higher scores for the $+$ explicitly defined words ($F(1, 73) = 7.29, p < .01$). These results are consistent with those of the preceding study and suggest that combining the data may obscure existing group differences. Both sets of results point to the importance of conducting a study specifically designed to determine which set of data are the best indicators of the latent knowledge monitoring variable.

The analysis of differences between high and low knowledge monitoring groups on predicted, postdicted, and actual AP scores, and final grades was also similar to that in the preceding study, with one addition—students’ actual final grades in the course were available as an additional dependent variable. Two groups were created by splitting students at the median on total number of accurate metacognitive estimates and computing a MANOVA to examine the significance of differences on the AP and grade data; nine students were eliminated due to missing information. No differences between groups who read either the text or the SVT were found ($F(5, 58) = 1.37$). Unlike the prior study, the differences between metacognitive groups only approached significance (Transformed Wilks lambda $F(5, 58) = 2.21, p = .066$; effect size = .16). Univariate analysis indicated that the high metacognitive group had significantly higher AP scores (Mean = 45.2) than the low group (Mean = 36.7; $F(1, 62) = 10.02, p = .01$); there were no differences on expected score either before or after the AP exam was taken, or on expected and actual final grades.

The findings that the high and low knowledge monitoring groups differed only on actual AP test performance, rather than on any of the estimates, also replicated those of the prior study. The failure to find differences on final grades may have been a function of the limited range of the grades; A–D grades (no F grades occurred in this sample) were converted to their numerical equivalents yielding only four scores. Furthermore, 76% of the grades were B or higher, further limiting their variability. The interaction between metacognitive group and those who read either the text or SVT was of borderline significance ($F(5, 58) = 2.18, p = .07$), probably principally attributable to the fact that the low knowledge monitoring group’s estimates of their AP scores and their final course grades were actually higher than that of the high monitoring group, although their actual scores and grades were lower than those of the other students.
An identical MANOVA was computed with students split at the median on the vocabulary test as the independent variable. There were no differences between groups who read either the text or the SVT. There were significant differences between the groups (Transformed Wilks lambda $F(5,58) = 5.70, p < .001$; effect size = .33). Univariate analysis indicated that the high vocabulary group also had higher AP scores (Mean = 47.0, $F(1,62) = 22.89, p < .001$) than the low group (Mean = 35.1). Unlike the analysis based on metacognitive estimates, the high vocabulary group also received higher final grades (Mean = 90.4, $F(1,62) = 5.24, p < .05$) than the low group (Mean = 85). The interaction between groups who read either the text or SVT and vocabulary groups were not significant (Transformed Wilks lambda $F(5,58) = 2.12, p = .076$) even though the lower vocabulary group predicted and postdicted higher AP scores and final grades and actually obtained lower scores on all three.

The second set of analyses examined the relationship between the knowledge monitoring scores and in-class student performance indices, such as the quizzes administered to students and scores on the essay and multiple-choice parts of the final examination. Because the instructor informed students that only the 10 highest scores on the 12 quizzes would count for the final grade, many students missed some quizzes. Therefore, for students taking at least 10 of the quizzes, the mean score on all the quizzes taken was used as one of the dependent variables. Students were split at the median on the correct knowledge monitoring scores, and MANOVA was computed on the quiz and final examination data; missing data limited this analysis to 70 students. No significant differences on class performance indices were obtained between the groups taking the SVT or reading the text $F(4,63) = 1.04)$. There was an overall significant difference between the high and low knowledge monitoring groups (Transformed Wilks lambda $F(3,64) = 4.36, p = .01$, effect size = .17). Univariate analyses indicated that the high knowledge monitoring group had significantly higher scores on the multiple-choice part of the final examination (Mean = 25.1) compared to the low group (Mean = 21.2, $F(1,66) = 12.66, p < .01$). Differences between the groups on the mean quiz score were not significant ($F(1,66) = 3.02, p = .09$), although the high knowledge monitoring group received higher scores (Mean = 4.51—each quiz had a total of six raw score points) than the low group (Mean = 4.1). There was no interaction between knowledge monitoring and whether groups read the text or not.

The identical analysis was computed with students split at the median on the number of words correct on the vocabulary test as the
independent variable. The high and low vocabulary score groups had overall differences (Transformed Wilks lambda $F(3,64) = 6.44$, $p <.01$, effect size = .232). The high vocabulary group had significantly higher scores on both essay (Mean = 17.2, $F(1,66) = 7.44$, $p <.01$) and multiple-choice (Mean = 25.5, $F(1,66) = 18.72$, $p <.001$) parts of the final exam, and on the mean of the quizzes (Mean = 4.6, $F(1,66) = 7.13$, $p = .01$) than the lower scoring groups (Means = 14.5, 20.9, and 4.0 respectively). In this study as in the preceding one the differences between vocabulary score groups were greater than the metacognitive estimates for differentiating students on the AP and final grade data (.33 effect size vs. .16) and on the class tests (.23 compared to .17 effect size).

Knowledge Monitoring and Performance Estimation Among Vocational High School Students

In Study VI, examining relationship between knowledge monitoring and school learning among vocational high school students, the participants were also asked to predict and postdict their grades on a course final examination; the actual score on that test was available as a dependent measure. MANOVAs indicated that neither the metacognitive knowledge monitoring estimates, nor the raw scores were significantly related to either of the dependent variables. The failure to find any differences on actual score is at variance with the findings of the two preceding studies using college students.

There are a number of differences between the studies using vocational high school and college students, in addition to the population differences, that may account for the diverse findings. The vocational high school students were asked to predict performance on a final exam in the class they were taking, and presumably had a much better idea of the content of the exam and how to prepare for it than the college psychology students, who had very little basis for knowing what to expect on the AP test, and could not prepare for it at all. Furthermore, because the vocational students had been graded on other exams in that class, they—unlike the college students—knew what grade to expect from their prior history in that class. These prior experiences may have been more important in determining the vocational high school students’ estimates than either their knowledge, or the metacognitive knowledge monitoring abilities.

Summary: Performance Estimation and Knowledge Monitoring

One striking finding of two of the studies using college students was that the strongest effects were found for students’ actual perfor-
mance, either on tests or in class, rather than for their estimates. Students' estimated performance on the AP exam, or their predicted class achievement, were typically not significantly related to KMA scores. On the other hand, performance on the AP test, or in class final exams (at least the multiple-choice part of the exam in Study VIII) were significantly related to knowledge monitoring. These results may be partially attributable to unrealistic estimates of students in the lower knowledge monitoring groups that were often higher (though not significantly so) than those of students in the high knowledge monitoring group.

There is a large difference between the accuracy of vocational high school students' test performance estimates and that of college students. The correlations between predicted and actual scores for the vocational students were .71, and .75 (p < .001) for postdiction; comparable results for college students in Study VII were .13 and .16, both nonsignificant, and for Study VIII they were -.14 and -.12, also nonsignificant. The greater accuracy of the vocational students is probably attributable to their familiarity with the material they were tested on, compared to the novelty of the AP test for the two college samples. As expected, the relationships were higher, though not significantly so, for postdiction, when students knew what was covered on the test, than for predictions confirming findings by Pressley and Ghatala (1990) who reported that students were generally more accurate in predicting their recall of text after completing a test than before taking it.

In both of the studies using college samples, the analysis of school performance data based on actual knowledge (number correct on the vocabulary test) accounted for more variance than comparable analyses using knowledge monitoring scores. It seems possible that students' achievement in classes is best predicted by actual knowledge, rather than estimates of it. Furthermore, in view of the nonsignificant relationships for the vocational high school sample between either knowledge or metacognitive estimates and class final exams, it seems likely that domain specific knowledge may be most useful for such predictions.

An important question to investigate is whether knowledge estimates in the domain in which school instruction and evaluation are likely to occur account for more variance than the actual knowledge, or than either estimates or knowledge of fairly general materials such as those used in these studies. The prior research assumed that the word list, vocabulary, and text passage were similar to the kind of material students would learn in nontechnical areas of school instruction. The studies relating knowledge monitoring to school learning found KMA relationships with achievement in English and Hu-
manities courses, but not for Science and Social Science. These results suggest that general knowledge, or metacognitive estimates of that knowledge are less useful in more technical areas, which rely on a domain specific technical vocabulary, than they are in subjects that have a more widely shared knowledge base and vocabulary.

**METACOGNITION AND MATHEMATICS**

All of the studies described so far used the KMA in the domain of declarative word knowledge, and employed similar or identical versions of the materials. A question arises whether the procedure generalized to domains other than vocabulary, such as mathematics. Like vocabulary, mathematics is of special interest because it is also important in school learning. In addition, however, the computation and problem solving in mathematics involve more procedural knowledge than does vocabulary learning. Thus, one purpose of the two studies described below was to examine the applicability of the knowledge monitoring procedure to the domain of procedural knowledge in mathematics.

The research described above used relatively mature students, predominantly those attending college; only two investigations used high school students. A further question examined in the next two studies was whether the KMA was equally useful with younger, elementary school students.

**Study IX. Monitoring Mathematical Problem Solving Among Elementary School Students**

Van Haneghan and Baker (1989) reported a number of investigations of the effects of metacognition on the accuracy of problem representation in mathematics. The results indicated that metacognition was as important for the learning of mathematics as it was for reading. These findings are supported by the expectations and results of other researchers, such as Campione, Brown, and Connell (1989), Lester, Garofalo, and Kroll (1989), as well as Schoenfeld (1992). Furthermore, research (Cardelle-Elawar, 1992; Montague, 1992) has also shown that students’ performance in solving mathematical problems was facilitated when they were instructed with a metacognitive approach. Therefore, it was expected that procedural KMA in mathematics should be related to general achievement in that subject, and to students’ ability to solve mathematical problems specifically.

The data for this study were collected by Dhalma Rosado. This investigation was presented as part of a paper at the annual convention of the American Educational Research Association, April 1995, San Francisco, CA.
Participants and Procedures

A list of 30 mathematical questions was constructed (20 computation, and 10 problem-solving items); the items were selected from the students’ fifth-grade mathematics curriculum. Students were first asked to take 6 minutes to determine if “you feel able solve these problems. Do not solve them now,” giving them an average of 12 seconds per problem. During a later session, the same 30 questions were readministered, and students were given 40 minutes to actually solve the problems. A number of anxiety scales were also administered.

A total of 51 fifth grade students (31 females) from an urban public school served as participants in this study. The students were predominantly of Hispanic origin, and their reading and mathematical achievement test scores ranged from average for their grade, to two years below grade level.

Results and Discussion

The scoring for the mathematics materials was similar to the vocabulary KMA. Four scores, like those used in the prior studies, were generated: Students felt that they could (a) solve a problem and did so (+ +); (b) not solve a problem and did not (- -); (c) solve a problem, but did not (+ -); and (d) not solve a problem, but did (- +). The results dealing with anxiety will be discussed later in the chapter.

There were no differences attributable to gender on students’ metacognitive estimates, so these data were pooled for further analysis. The knowledge monitoring scores were correlated with the total math score on the Metropolitan Achievement Test (1985) obtained from the students’ records. The correlations are displayed in Table 4. The last row in that table represents the number correct on the math test. The + + and - -scores were combined to indicate correct estimates of students’ ability to solve mathematical problems, and the - + and + - scores were added to form the incorrect estimates.

Table 4 indicates that three of the four estimates were significantly related to students’ mathematics achievement. The correlation between number correct on the math test and Metropolitan score was .52. When that relationship is compared to the correlation of .73 between Metropolitan score and + +, or the correlation of .76 between the Metropolitan score and total number of correct estimates, it is clear that metacognitive estimates of the ability to answer the questions are more substantially related to mathematical achievement than the number of problems solved correctly, irrespective of esti-
Table 4. Correlations Between Different Knowledge Monitoring Scores and Achievement in Mathematics.

<table>
<thead>
<tr>
<th>Score</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>++</td>
<td>.73***</td>
</tr>
<tr>
<td>+ -</td>
<td>-.43**</td>
</tr>
<tr>
<td>- +</td>
<td>-.65***</td>
</tr>
<tr>
<td>- -</td>
<td>-.11</td>
</tr>
<tr>
<td>++ &amp; -</td>
<td>.76***</td>
</tr>
<tr>
<td>- + &amp; + -</td>
<td>-.72</td>
</tr>
<tr>
<td>No. Correct</td>
<td>.52</td>
</tr>
</tbody>
</table>

dicate. That finding was confirmed by regression analysis. When the number of correct estimates, incorrect estimates, and total number right were in the model, only the correct estimates contributed significantly to the prediction of Metropolitan score (R\(^2\) Change = .08, F(3,45) = 8.52, P < .01). These results confirm the basic assumption of the knowledge monitoring procedure that students’ metacognitive judgments contribute significant independent variance beyond those accounted for by number correct on a test.

The results support predictions regarding the relationships between the procedural KMA in mathematics and achievement in that domain. As expected, there were substantial correlations between students’ estimates of their ability to solve mathematical problems and their achievement in mathematics. Also as expected, inaccurate assessments were negatively related to achievement. Although no causal inferences about mathematical achievement and knowledge monitoring can be made from these correlational data, the fact that the variables covary as expected supports the generalizability of the procedure and suggests that the technique is useful for further research in mathematics.

Study X: Relationship of Procedural KMA in Mathematics with Age and Achievement\(^8\)

The prior study provided encouraging evidence of the knowledge monitoring procedure’s applicability to procedural knowledge in mathematics. Furthermore, the results of Study IX also indicated that

\(^8\)The data for this study were collected by Audrey D’Agostino. The study was part of a paper presented at the annual convention of the American Educational Research Association, April 1994, in New Orleans, LA.
the KMA could be used with elementary school students. Because metacognition is often viewed as a developed ability and assumed to increase with age, one purpose of the next study was to investigate whether procedural knowledge monitoring in mathematics would also increase with age. The preceding study indicated a high relationship between KMA scores in mathematics and achievement test scores in that domain. Study X also examined whether knowledge monitoring scores were related to teachers' judgments of mathematical ability.

Participants and Procedures

Students were presented with 15 mathematical word problems involving addition and subtraction. The problems were set in the context of an ice cream store and students received a menu of prices for different products referred to by the word problems. The materials were prepared in two versions presumed to elicit varying interest levels among students. The results dealing with interest are discussed later in the chapter. The materials were administered on two days during regular class periods. On the first day, students examined the

![Figure 7. Mathematical Knowledge Monitoring Scores by Grade Level.](image-url)
problems and estimated whether they could solve them or not; on the second day the students were asked to solve the problems.

Students (N = 164, 70 female) were selected from the fourth, fifth, and sixth grades of a school attended largely by children from minority groups. Mathematical ability was determined by teachers' judgment; 59 students were classified as being in the low, 67 in the medium, and 81 in the high ability groups.

Results and Discussion

Students' responses were assigned a score of 1 for correct estimates (combining the ++ and -- scores), and 0 for incorrect estimates (combining the + - and - + scores). Due to a computer malfunction, the raw data were not available for rescoring to form the same scores used in the other studies. The data were then submitted to a 3 (grades) x 2 (sex) x 2 (group) x 3 (math ability) analysis of variance. As indicated above, the results dealing with interest are discussed later.

As expected, a significant increase in knowledge monitoring scores from grades 4 to 6 was found ($F = 34.66, df = 2, 144, p < .001, \eta^2 = .26$; see Figure 7 for plot of the data). Also as expected, the results indicated that knowledge monitoring scores increased with math-

![Knowledge Monitoring Scores](image)

Figure 8. Mathematical Knowledge Monitoring Scores by Math Achievement Group.
ematical ability ($F = 15.25$, $df = 2$, $144$, $p < .001$, $\eta^2 = .18$; see Figure 8 for plot of the data). These results offer further support for the construct validity of the knowledge monitoring procedure because older or more capable students were expected to have better metacognition than their younger, less capable counterparts. There were no significant differences attributable to sex.

Summary: Knowledge Monitoring and Mathematics

The results of the two studies using the procedural KMA in mathematics were quite positive regarding its applicability to that domain. The relationship of monitoring scores to achievement in mathematics in Study X are similar to the correlations with math achievement test scores reported in Study IX, and both indicate strong relationships between metacognitive knowledge monitoring and achievement in mathematics. The increases in metacognitive ability associated with age reported in Study X also support that relationship. Furthermore, 10 of the studies reviewed in this chapter examined students’ estimates of their declarative word knowledge. Because most of the items in both mathematical studies were composed of procedural knowledge needed to solve word problems, the results suggest that the KMA may be applicable to that type of knowledge as well.

METACOGNITION AND AFFECT

The paradigm shift to a cognitive orientation in psychology generated a great deal of research to clarify the cognitive processes controlling learning from instruction. Although that work has identified many cognitive processes that are important in human learning, the impact of affective processes on such learning has received considerably less attention (Tobias, 1992, 1994a, b). The aim of the research discussed in this section is to forge a link between affect and cognition by examining the influence of affective variables such as anxiety and interest on metacognitive knowledge monitoring.

The Impact of Anxiety on Knowledge Monitoring

One of the affective variables that has been the subject of a great deal of research, both in educationally relevant situations and in others, has been anxiety and its impact on learning. In general, that research has suggested a negative relationship between different forms of anxiety and learning from instruction (Tobias, 1992; Hembree, 1988). It has been suggested (Tobias, 1985, 1992) that interference in
students’ performance as a result of anxiety was attributable to reduced cognitive capacity available for task solution. It was reasoned that the central representation of anxiety absorbs some proportion of cognitive capacity, leaving a reduced amount available for allocation to work on tasks. The further absorption of capacity required by an executive process such as metacognitive knowledge monitoring was expected to be especially debilitating for highly anxious students whose cognitive capacity is expected to be reduced by students’ concerns about their test anxiety. Therefore, a negative relationship between anxiety and knowledge monitoring was anticipated because “highly test anxious students can be expected to have less adequate metacognitive abilities than those with lower anxiety” (Tobias, 1992, p. 28).

Knowledge Monitoring, Reading Comprehension, and Test Anxiety

It will be recalled that Study II examined the relationship of the knowledge monitoring procedure to anxiety, in addition to reading comprehension. The worry subscale of the Test Anxiety Inventory (Spielberger et al., 1980) was administered to 117 undergraduates from a large urban university; 65% were women.

As expected, the more highly anxious participants also performed less well on the KMA. Those lower in anxiety achieved a significantly higher number of “hits” than those prone to higher levels of anxious worry ($t(115) = 4.92, p < .001$), and in general the less anxious had higher levels of metacognitive word knowledge as measured by $d'$ (multiple $r$ squared ($t(115) = 4.07, p < .001$), confirming the expected negative relationships between knowledge monitoring and test anxiety.

Knowledge Monitoring in Mathematics and Anxiety

Study II found the expected negative relationship between knowledge monitoring procedure scores and anxiety in the vocabulary domain. Study IX, in addition to investigating the extension of the knowledge monitoring procedure to mathematics, also studied its relationship with both test and mathematics anxiety.

As part of Study IX the Fenema-Sherman (1976) scales assessing math anxiety and attitudes towards mathematics were administered to the 51 participants (see the earlier description of Study IX) in a first session. In order to assure that the participating elementary school students could understand the questions, each item was read aloud as students read the materials. The Worry-Emotionality scale (Morris,
Davis, & Hutchings, 1981), a 10-item Likert-type measure of these components of state test anxiety, was also administered. Students' mathematical achievement was determined from their scores on the Metropolitan Achievement Test (1985) obtained from school files.

In Study IX no sex differences in the effects of anxiety were found, so the data for all students were pooled. The relationships between knowledge monitoring and mathematics anxiety (scored in the direction of higher anxiety yielding higher scores) and attitudes towards mathematics, as well as with worry and emotionality are shown in Table 5.

Table 5. Correlations Between Knowledge Monitoring Scores and Anxiety in Mathematics.

<table>
<thead>
<tr>
<th>Score</th>
<th>Math Anxiety</th>
<th>Worry and Emotionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>++</td>
<td>-.42**</td>
<td>-.22</td>
</tr>
<tr>
<td>+ -</td>
<td>.32*</td>
<td>.25</td>
</tr>
<tr>
<td>- +</td>
<td>.38**</td>
<td>.23</td>
</tr>
<tr>
<td>- -</td>
<td>.00</td>
<td>.20</td>
</tr>
<tr>
<td>++ and - -</td>
<td>-.46**</td>
<td>-.15</td>
</tr>
<tr>
<td>- + and + -</td>
<td>.46**</td>
<td>.33*</td>
</tr>
</tbody>
</table>

* = $p < .05$
** = $p < .01$

Table 5 indicates that, as expected, mathematics anxiety was negatively related to incorrect metacognitive estimates and positively related to correct ones. The correlation between number right and math anxiety was -.25 and not significant, though the relationships with metacognitive estimates were negatively and significantly related to test anxiety. The negative relationships between metacognition and anxiety are generally similar to those found in Study II, confirming expectations that anxious students have lower metacognition than their less anxious colleagues.

Anxiety and Knowledge Monitoring Among High School Dropouts and Continuing Students

Study VI investigated whether knowledge monitoring differed between continuing students and high school dropouts. An addi-
4. METACOGNITIVE KNOWLEDGE MONITORING

The purpose of that study was to examine the differences in anxiety between high school dropouts and continuing students, as well as the relationship between anxiety and metacognitive knowledge monitoring. In this study, the Test Anxiety Inventory (Spielberger et al., 1980) was given to all participants, followed by two administrations of the Worry-Emotionality Scale (Morris et al., 1981). Initially, participants completed the Worry-Emotionality scale in terms of the way they felt while being tested in general; when the scale was re-administered after the vocabulary test, students were asked to respond the way they felt while completing the vocabulary test. It will be recalled that the dropout group consisted of 42 individuals (14 female) who had dropped out of high school and the 47 (16 female) continuing students who had a school GPA of at least B-.

Surprisingly, the results of a MANOVA indicated that there were no anxiety differences between high school dropouts and continuing students on any of the seven anxiety scores (the three Test Anxiety Inventory scores: Worry, Emotionality, and Total, in addition to four Worry and Emotionality scores from each administration of those scales). That finding is puzzling in view of the reports in the literature that poor performance in school, and presumably on tests, was a major reason for students dropping out of high school. One reason for these findings may rest with the problems to which self-report measures in general, and self-reports of test anxiety in particular, are subject. It is easily possible for students to minimize or deny responses indicative of test anxiety on these measures, and to present themselves as not caring about how well they might function on tests. The knowledge monitoring procedure, however, made it difficult for students to present themselves in a more favorable light, and that may account for the findings of group differences in metacognitive knowledge monitoring and their absence in test anxiety.

Most of the zero order correlations between KMA scores and anxiety indices were negative, and a fair number were significant. Multiple linear regression analyses were computed with the KMA scores as the dependent variable, and the anxiety scores as the independent variable. Results indicated that the anxiety scales had a significant impact only on the + + scores, $R^2 = .25$, $F(7,72) = 3.43$, $p<.01$; significant beta weights were found for Emotionality, on the Worry-Emotionality Scale taken after students completed the vocabulary test ($t = 2.74$). The regression analysis also indicated that none of the other KMA scores were significantly related to the anxiety scales. In view of the number of anxiety and knowledge monitoring scores, the findings of significant relationships for some of them is not
surprising. In general, however, the results of this study suggested that there was little association between metacognitive knowledge monitoring and anxiety.

Anxiety and Knowledge Monitoring Among Vocational High School Students

A further purpose of Study V, examining metacognition among Vocational High School students, was to study the relationship between anxiety and metacognitive knowledge monitoring, as well as between anxiety and achievement. In addition to the rationale relating metacognition to anxiety, it was also expected that students with lower GPAs should have higher anxiety than those who learned more effectively. These two questions were examined in this study.

Recall that the anxiety scales, and the order in which they were administered in Study V, were identical to those employed in the study of high school dropouts (Study VI). The Worry-Emotionality scale (Morris et al., 1981) was administered first and students were asked to respond in terms of the way they felt while taking tests in general. The Test Anxiety Inventory (Spielberger et al., 1980) was then given, followed by a second administration of the Worry-Emotionality scale with instructions for students to respond the way they felt while taking the vocabulary test. A total of 61 students (59 male) participated in this study.

The significance of differences in anxiety scores between the participants in Study V above and below the median GPA in vocational high school was examined by computing a MANOVA. Surprisingly, there were no differences between the GPA groups on any of the seven anxiety scores. Also, much as in Study VI, most of the zero order correlations between knowledge monitoring and anxiety were negative. Multiple linear regression analysis was then computed with the metacognitive knowledge monitoring scores as the dependent variable, and the anxiety scores as the dependent variable. None of the regression equations were significant for this sample.

Summary: Anxiety and Metacognitive Knowledge Monitoring

The evidence regarding the relationship between anxiety and metacognitive knowledge monitoring is mixed. Significant negative relationships were expected and found in two of the studies, one in mathematics and the other using vocabulary. On the other hand, two other studies failed to find any evidence of differences. Study II, in which significant negative relationships with anxiety using the vo-
cabulary materials were found, had a much larger sample than the studies using vocational high school students (see Study V) or high school dropouts (see Study VI). Because many of the test anxiety-metacognitive knowledge monitoring relationships in the two latter studies were, as expected, in the negative direction, and because some of the regression analyses between these variables approached significance, further research with larger samples is clearly needed to clarify the relationship between anxiety and knowledge monitoring. The results of Study II suggest that metacognitive word knowledge and test anxiety each contributed to performance on less challenging reading. On demanding material, however, test anxiety and metacognitive knowledge monitoring ability appear to interact to affect performance. The highly anxious examinee, regardless of metacognitive ability, performed less well on the more demanding reading tasks, suggesting that anxious worrying can interfere with strategic use of metacognitive skill when the performance tasks are cognitively demanding. That finding is in accord with the anxiety-cognitive capacity model (Tobias, 1992), because more demanding tasks require greater proportions of cognitive capacity that may not be available as a result of the resources absorbed by anxiety. Further research is required to pursue that intriguing finding.

In Studies V and VI, the failure of a number of anxiety indices to differentiate between either high school dropouts and continuing students, or between students above and below the median in GPA was surprising. A meta-analysis of 562 studies dealing with test anxiety (Hembree, 1988) had indicated that lower achieving students had higher test anxiety than their more capable counterparts. Although there had been no prior research specifically relating test anxiety to dropping out of high school, the bulk of this literature has indicated that the concern of students about their academic achievement was a major factor in dropping out of school, clearly suggesting that differences in test anxiety could be expected. As mentioned above, the fact that both the studies dealing with dropouts and vocational high school students found significant knowledge monitoring differences in the expected direction, and neither found differences on a group of seven test anxiety scales re-emphasizes some of the problems with self-report measures described at the beginning of this chapter.

Although the nonsignificant results for anxiety in Studies V and VI may be attributable to small samples, or to other unknown factors, it should also be noted that the tendency of participants to present
themselves in a more positive light may well have contributed to the nonsignificant findings. One advantage of the KMA is that, because students do not report on either their feelings or their cognitive processes, it is difficult for them to present themselves more favorably. Of course, students could easily claim to know more words than they actually do. However, that claim is immediately checked by the administration of the vocabulary test making it harder for students to appear in a more positive light.

Knowledge Monitoring and Interest

There has been a good deal of recent research on the effects of interest on learning (Renninger, Hidi, & Krapp, 1992). One reason for that increase is suggested in one definition: “intrinsically motivated behaviors are those the person undertakes out of interest” (Deci & Ryan, 1991, p. 241). From that perspective, clarification of the effects of interest also adds to an understanding of the impact of intrinsic motivation on learning. Second, interests appear stable and long lasting among adults (Hidi, 1990; Schiefele, 1991), suggesting that instruction adapted to students’ interests may have positive motivational characteristics for long periods of time. Third, interests are ubiquitous—everyone is interested in something. Fourth, findings of surprisingly variable and ineffective cognitive processing of instruction (Paris, 1988; Tobias, 1989) suggested that these results may be attributable to the possibility that students’ interests or motivation were not engaged by the materials used in such studies. Finally, research on interest provides a useful and educationally relevant avenue for study of the relationship between affect and cognition (Tobias, 1989, 1994a, b)—a much needed clarification in order to obtain a more complete picture of people’s day to day functioning.

Research has indicated that reading comprehension and recall are facilitated when students work on material related to their interests (Renninger et al., 1992). Furthermore, Schiefele (1990, 1991, 1992a, b) found that comprehension of interesting text was “deeper” (i.e., more likely to be propositional than verbatim). Little is known, however, about the cognitive processes that mediate the effect of interest on comprehension and recall of reading. Therefore, it was recommended (Renninger et al., 1992; Tobias, 1994a) that research concentrate on the identification of the processes invoked by interest to facilitate learning. The purposes of the studies reported in this section of the chapter are to examine whether interest improved students’ metacognitive knowledge monitoring.
Situational Interest, Topic Interest, and Knowledge Monitoring

Two types of interest, *situational* and *topic*, have been distinguished (Renninger, Hidi, & Krapp, 1992). *Situational interest* is elicited by aspects of a situation, such as its novelty or intensity, and by the presence of human interest factors contributing to the attractiveness of different types of content. *Topic interest* refers to peoples' relatively enduring preferences for different topics, tasks, or contexts and how they influence learning. The effects of both types of interest on metacognitive knowledge monitoring were investigated in this study. It was expected that students with greater topic interest, and text that elicited situational interest, would lead to more accurate knowledge monitoring. Furthermore, because interest has been found to lead to deeper types of text processing (Schiefele, 1990, 1991, 1992a, b), it was expected that students would make more accurate knowledge monitoring judgments on words requiring intense processing if the material were interesting rather than neutral. The meanings of implicitly defined words must be inferred, whereas those defined explicitly merely require recall of the definitions. Therefore, it was reasoned that the meanings of implicitly defined words should be estimated more accurately on interesting content compared to more neutral content.

Study III Revisited

Recall that two groups of students were used in Study III, nursing students and college freshmen. Because the passage dealt with heart disease, it was expected that nursing students would have greater topic interest in that material than would the freshmen. Situational interest was varied by converting the expository passage to a narrative format. The narrative passage contained story attributes, such as character identification and life themes, which according to Hidi and Anderson (1992), increased the situational interest of passages. A principal character was introduced in the narrative version, which then described his efforts to learn more about coronary disease because his father had developed a mild form of that illness. The passage indicated that he was trying to help his father prevent the development of more serious coronary problems. This structure made it possible to include in the narrative version all the factual information present in the expository version of the passage. Only 84

---

This study conducted by Sigmund Tobias was published in the *Journal of Educational Psychology*, 1995. See References.
of the 139 students completed all the materials during two sessions. Complete data were available for 33 nursing students and 51 freshmen.

In Study III the correct metacognitive judgments (combining ++ and -- scores) were submitted to analysis of variance, with the correct estimates on explicitly and implicitly defined words—the dependent variables—treated as a repeated measure. In view of the importance of controlling for prior knowledge differences in interest research (Tobias, 1994a), students’ scores on the first administration of the vocabulary test were used as a covariate because the nursing students were more familiar with the heart disease material (Prescore Mean = 27.4, SD = 4.0) than were the freshmen (Prescore Mean = 20.1, SD = 5.3). Because there were an unequal number of females in the groups (24 of 51 freshmen and 28 of 33 nursing students), gender was added as a factor in the design. Thus, the ANOVA consisted of a full 2 (freshmen vs. nursing students) x 2 (expository vs. narrative passages) x 2 (gender) factorial design, with prescore as a covariate. Again, the two-level repeated measure consisted of the number of correct judgments on explicitly and implicitly defined words after reading the text. The main effect of the repeated measure was assessed in the “deviation” manner described by Delaney and Maxwell (1981).

The ANOVA results indicated that there was a significant overall difference between the freshmen and nurses ($F(1,75) = 4.99, p < .05$), favoring nursing students. In addition, the mean number of correct estimates was higher for explicitly than for implicitly defined words ($F(1,75) = 8.27, p < .01$). None of the other main effects or interactions was significant. The covariate, number correct on the first administration of the vocabulary test, exerted a significant effect on the dependent measures ($F(1,75) = 17.01, p < .001$). The adjusted means for freshmen on correct estimates for explicitly and implicitly defined words were 13.7 and 12.5, respectively, and for nursing students the corresponding means were 15.0 and 14.1. Ideally, future research should use participants with similar prescores who differ in their interest in such medically relevant materials.

These results support the general hypothesis of enhanced metacognition for topic interest. As anticipated, nursing students, for whom the heart disease passage was more interesting than for freshmen, made more accurate metacognitive assessments of their vocabulary knowledge than the freshmen, even when differences in their prior knowledge of the vocabulary was controlled statistically. The expected differences attributable to situational interest were not found
because the narrative and expository passages resulted in similar KMA scores. Finally, contrary to expectations, explicitly defined words were judged more accurately than those that were implicitly defined for both nursing students and freshmen.

The absence of knowledge monitoring differences due to situational interest may be a function of the similarities between the expository and narrative texts. Even though the passage was altered to elicit differences in situational interest, ratings of interest on a Likert-type scale, in the original study and on a follow-up, failed to find any differences between the passages. Perhaps, greater differences in content are needed to result in situational interest differences than occurred in Study III.

Interest and Knowledge Monitoring in Mathematics Among Elementary School Students

Study X found that metacognitive knowledge monitoring ability in mathematics increased with grade and mathematical ability. A further purpose of that study was to examine the impact of personalizing instruction on metacognition. Research (Anand & Ross, 1987; Bracken, 1982; Herndon, 1987; Lopez, 1999, 1990; Ross & Anand, 1987; Wright & Wright, 1986) has shown that personalizing mathematical word problems by including materials such as the names of students, their friends' and teachers' names, or including materials related to their interests improved performance and attitudes to the materials. These, and similar, studies suggested that heightened interest was aroused by personalizing word problems. It was, therefore, hypothesized that the elevated interest should improve students' metacognitive knowledge monitoring.

Participants in Study X (N = 164, 70 females, and all of whom were selected from fourth, fifth, and sixth grades of a school attended largely by children from minority groups) were randomly assigned to either interesting or control materials. In the interesting condition the names of classmates and teachers were included in word problems, whereas the materials used for the control group used standard rather than familiar names. In each condition, 15 mathematical word problems set in the context of an ice cream store were presented and students received a menu of prices for different products and were required to add and subtract menu items. A 12-item Likert-type scale designed to assess interest in the materials was also administered.

In this study students' responses were assigned a score of 1 for correct estimates and 0 for incorrect estimates. The data were then submitted to a 3 (grades) x 2 (sex) x 2 (interest group) x 3 (math ability)
analysis of variance. The findings dealing with knowledge monitoring and mathematical ability and grade level were reported previously. There were no significant differences attributable to sex or to interest. However, there was an interaction between math achievement level, as determined by teacher judgment, and interest ($F = 6.02$, $df = 2, 144$, $p < .01$, $\eta^2 = .05$; see Figure 9 for a plot of the data).

The interaction, unlike the main effect found in the prior interest study, suggests that the personalization improved the performance of low ability math students, but had little effect on the other groups. It seems possible that setting the word problems in the context of an ice cream store may have raised the interest level of the materials for both groups, thus leading to the insignificant main effect for interest. In view of the known difficulties students have with math word problems (NAEP, 1979), it was thought to be important to make the materials interesting for both groups by situating them in an ice cream parlor. There is evidence that this setting did arouse the interests of all students. There were no differences ($F < 1.0$) between the high and low interest groups on the 12-item Likert-type scale administered after

![Figure 9. Mathematics Knowledge Monitoring Scores by Interest Group.](image_url)
students completed the problem solving. Furthermore, there were also no differences between the high and low interest groups in the number of problems solved correctly. These findings indicate that even the low interest group may have found the materials more attractive than the math word problems usually received in school, and suggests that an overall facilitative effect for interest may be found when the materials elicit greater differences in interest between the groups.

**Summary: Affect and Knowledge Monitoring**

The findings of the anxiety and interest studies indicate that anxiety generally seems to have a negative effect on metacognitive knowledge monitoring, and that working on interesting materials seems to facilitate it. Further research is needed to answer many questions before these tentative conclusions can be stated with greater confidence. It seems, however, that the knowledge monitoring procedure is a useful way for studying the effects of affect on metacognition, and especially of investigating the effects of interest. There are a number of persuasive models specifying the cognitive processes mediating the impact of anxiety on learning (Sarason, 1987; Eysenck, 1988; Tobias, 1992). However, little is known about the cognitive processes by which such “positive” affective variables as interest and motivation facilitate learning. The knowledge monitoring procedure seems useful for further research relating metacognition to such positive variables as interest or intrinsic motivation.

**METACOGNITIVE KNOWLEDGE MONITORING AND OTHER VARIABLES**

Most of the studies reported above related the knowledge monitoring procedure to relatively traditional variables such as achievement in school, estimates of achievement, anxiety, and interest. Two studies examined the relationship of the knowledge monitoring procedure to other variables such as need for feedback and the procedure’s ability to differentiate between different types of students. These studies are summarized below.

**Study XI. Knowledge Monitoring Procedure and Need for Feedback**

Feedback or reinforcement is one of the best known variables in learning research. Numerous studies have demonstrated that feedback facilitates learning. McKeachie (1974) suggested that the effects

---

10The data for this study were collected by Nadia Seignon.
of feedback or reinforcement on human learning are not uniform, but may vary with individuals and situations. Ashford and Cummings (1983) found that the importance of feedback varied with individuals’ uncertainty and Tuckman and Sexton (1992) found that students in a no-feedback condition who held high beliefs in their own performance capability outperformed those receiving feedback, whereas the reverse was true for students of middle and low self-perceived capability. These results clearly supported the idea that there were individual differences in the need for feedback.

It was expected that the need for feedback ought to depend on students’ metacognitive capability to monitor their knowledge gathering activities. In an analysis similar to that proposed by Butler and Winne (1995), it was reasoned that students with accurate knowledge monitoring abilities probably rely more frequently on their own, or internal, feedback regarding the accuracy of their responses than on their less accurate peers. Such students are likely to have learned from experience that external feedback often duplicates the information supplied internally, and should require less externally supplied feedback than colleagues with less accurate knowledge monitoring abilities. Therefore, when students have a choice of whether they choose to obtain feedback or not, a negative relationship between KMA accuracy scores and amount of feedback was expected.

Participants and Procedures

A list of 25 words, appropriate for fifth grade students, and a vocabulary test based on the same words were developed. Participants were also given a reading test consisting of 11 narrative stories with an average length of 140 words, or 15 sentences. Each story had a blank to be filled in, and students were instructed to select a word from four choices appearing in the right margin for each blank; the words on the reading test and word list were different. Participants were told that the correct answer to each question was printed in the left margin of each page, covered by a tab, and they could look at the answers whenever they wished simply by lifting the tab. Participants were tested individually, and the number of times the tabs were lifted to inspect the correct answer were recorded.

A sample of 59 fifth grade students (35 females) participated in this study. The school was attended primarily by minority students.

Results and Discussion

Students’ need for feedback was operationally defined as the number of times they lifted the tabs covering the correct answers to
questions in order to check on the correctness of their response. The knowledge monitoring procedure was scored to determine students’ accuracy in estimating their word knowledge and the results were then correlated with amount of feedback sought on the reading test. The results of that analysis are show in Table 6.

As expected, the results indicated that amount of feedback was substantially related to the ability to monitor knowledge accurately.

Table 6. Correlations with Amount of Feedback.

<table>
<thead>
<tr>
<th>Score</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>++</td>
<td>-.50**</td>
</tr>
<tr>
<td>+ -</td>
<td>.38*</td>
</tr>
<tr>
<td>- +</td>
<td>.56**</td>
</tr>
<tr>
<td>- -</td>
<td>-.13</td>
</tr>
<tr>
<td>++ &amp; - -</td>
<td>-.79</td>
</tr>
<tr>
<td>- + &amp; + -</td>
<td>.76</td>
</tr>
<tr>
<td>Score (++ &amp; - +)</td>
<td>-.19</td>
</tr>
<tr>
<td>R</td>
<td>.84</td>
</tr>
<tr>
<td>R²</td>
<td>.71</td>
</tr>
</tbody>
</table>

The accuracy of knowledge monitoring was substantially and negatively related to amount of feedback ($r = -.79, p < .001$), as were the number of inaccurate estimates ($r = .76, p < .001$). Equally interesting was the finding that vocabulary knowledge, determined by the number correct on the vocabulary test, was not related to amount of feedback ($r = -.19$). The findings suggest that, as expected, need for feedback was heavily related to the ability to monitor one’s knowledge accurately. Furthermore, estimates of students’ knowledge were clearly the major contributor to that relationship, because actual knowledge was unrelated to amount of feedback.

An equally important aspect of this study and its results was the fact that a new word list and vocabulary test was developed, different from the materials used in any of the other studies described in this chapter. Therefore, the findings also indicated that the knowledge monitoring procedure has some generality across different types of vocabulary materials. Furthermore, this was the first study using a declarative vocabulary KMA with elementary school students, and the results suggest that the procedure was as applicable to younger students as were the mathematical materials used in Studies IX and X.
Study XII. Knowledge Monitoring Differences Among Learning Disabled and Hyperactive Students¹¹

It has been shown (Brown & Campione, 1986; Swanson & Trahan, 1992) that students diagnosed as learning disabled (LD) have lower metacognition than those who are diagnosed as not being LD. Similarly, students with Attention Deficit Hyperactivity Disorders (ADHD) have been succinctly described by Douglas, Barr, O'Neil, and Britton (1986) as having a cognitive deficit featured by an inability to stop, look, listen, and think. A review of research dealing with ADHD (Westby & Cutler, 1994) indicates that such students tend to have less effective complex problem-solving strategies and organizational skills, that they use less efficient strategies on memory tasks, that they "demonstrated deficits on all measures of study behavior. They studied for less time, expended less effort, and used poorer strategies...students with ADHD have significant deficits in executive processes" (Westby & Cutler, 1994, pp. 63-64). These deficits clearly suggest that ADHD students have less effective metacognition. Therefore, students diagnosed as LD, or ADHD, should have less accurate knowledge monitoring capabilities than students not affected by these conditions. It was the purpose of this study to test that hypothesis.

Participants and Procedures

A list of 35 words, and a vocabulary test based on the same words, were developed from the high school curriculum. Participants (N = 90) were selected from the ninth (N = 29) and tenth (N = 61) grades of a public high school in an urban area; there were 28 females and 62 males. LD and ADHD groups (N = 30 each) were formed by selecting students diagnosed by a school-based support team consisting of an educational evaluator, a school psychologist, and a social worker; scores on the Degrees of Reading Power (Touchstone, 1991) test placed these groups in the 15th percentile of the population. A contrasting student group (N = 30) was selected on the basis of having average reading ability on the DRP and no histories of special educational needs.

Results and Discussion

Three of the KMA scores (+ +, + - , and - -) were analyzed by MANOVA (the fourth score [- +] could not be entered due to linear

¹¹The data for this study were collected by Julie Wilson.
dependencies), with sex and group as the independent variables. A significant overall difference among the groups was found (Transformed Wilks lambda $F(6,164) = 5.95, p<.001$, effect size = .179). Univariate analyses indicated significant differences among the groups on ++ scores ($F(2,84) = 16.02, p < .001$; Control Group Mean = 28.4; LD Mean = 22.2; and ADHD Mean = 23.0). Univariate analyses also indicated another difference on the - - score ($F(2,84) = 5.32, p < .01$; Control Group Mean = 1.5; LD Mean = 3.6; and ADHD Mean = 4.3); students in the control group had lower scores because they had fewer incorrect answers. There were no differences attributable to sex, and no interaction between sex and group was found.

A similar analysis of the number correct on the vocabulary test (+ + and - +) also indicated significant group differences ($F(4,166) = 7.55, p < .001$, effect size = .154). Univariate analysis indicated that only the differences on the ++ scores were significant; the group means are the same as for the preceding analysis. The results confirm expectations regarding differences between regular, LD, and ADHD students with respect to their ability to monitor their knowledge and differentiate between what they know and do not know in this domain. Although the results were similar when the dependent variable consisted only of the number correct on the vocabulary, the effect size on the latter analysis was smaller (.154 compared to .179). As expected, the control group of students without special needs were more able to differentiate between the ++ and - - words than students in the other two groups.

There were large differences in reading ability between the groups which may also have accounted for the group differences, irrespective of diagnostic category. It is often difficult to separate the effects of reading ability from research comparing LD, ADHD, and more traditional students because the presence of reading problems is one of the defining characteristics of the two former groups. Further research with more similar groups may clarify this problem. In any event, these results provide additional support for the construct validity of the metacognitive knowledge monitoring procedure. In view of the fact that this study, like the prior one, also developed a new list of words and vocabulary test, the results also support the generality of the procedure across different types of content.

GENERAL DISCUSSION

The findings of the 12 studies summarized above support the construct validity of the KMA. Comparable results were found for samples from student populations such as students in elementary
school, students attending regular and vocational high schools (including students diagnosed as LD and ADHD), individuals who dropped out of high school, and students in pre-college-admission status, and students in their second or higher year in college. Furthermore, substantially similar results were obtained for procedural knowledge in mathematics, in addition to declarative vocabulary knowledge based on three different vocabulary sets developed to be appropriate for elementary school through college levels.

The results suggest that the procedure has some generality across different populations, declarative and procedural knowledge, as well as different types of vocabulary. In view of the fact that the KMA may be group administered and/or given by computer, and is objectively scored, it seems to be a promising approach for the assessment of the knowledge monitoring component of metacognition. Furthermore, Studies V and VI indicated that the KMA made it less likely that students presented themselves in a more favorable light than self-report scales of anxiety, one of the problems inherent in self-report instruments. Although no data comparing the KMA to other metacognitive scales have so far been collected, we expect that this measure of knowledge monitoring is likely to be more accurate than self-report scales because students are less able to present themselves in socially desirable ways. It remains for further research to investigate this possibility.

The KMA’s relationships with external criteria were somewhat variable. Relationships with standardized achievement tests were substantial and significant. For example, in Study I correlations with a reading comprehension test were .67. Similarly, relationships with achievement in mathematics were also substantial in Study IX (r = .76) and in Study X, significant effects were found for KMA differences in students’ math achievement (eta² = .26) and for higher levels of mathematical performance across three elementary school grades (eta² = .18). Pintrich et al. (this volume) cite some of these findings as being among the most positive relationships between any metacognitive measure and external criteria. Relationships with need for feedback in Study IX were also found to be substantial (r = .62). Significant, though somewhat more moderate, relationships were found in studies in which the KMA differentiated between known groups such as regular students and dropouts (Study VI), or among LD, ADHD, and students without special needs (Study XII). Generally the lowest, though frequently significant, relationships were found between KMA scores and college grades. Presumably, as indicated previously, the low reliability of such grades accounts for the modest associations with
grades. It should also be noted that differences between the effects of knowledge estimates and actual knowledge, discussed below, should be considered in examining the effects.

A number of issues raised by the results require further research. These include such questions as: Do multiple administrations of the knowledge monitoring procedure increase its relationship with other variables? Which of the different scores are optimal indicators of knowledge monitoring abilities? Do estimates of knowledge account for more variance than the actual knowledge? These questions are addressed below.

KMA and Dynamic Assessment

Some of the studies described above administered the text passage to only a part of the sample, others did not use the text passage at all, and still others gave a word list and vocabulary test before and after students read a text passage from which the word meanings could be inferred. A question arises about the value of interspersing the text passage between administrations of the word list and vocabulary. Giving students a chance to update their knowledge has some similarities to dynamic assessment approaches (see Carlson & Wiedl, 1992; Guthke, 1992; Lidz, 1992) in which students are given the opportunity for new learning before being tested. Dynamic assessment procedures usually also include some intervention in students' attempts to learn, observations of their reaction to the intervention, and an evaluation of students' responses to the assistance as part of the assessment. Reviews have suggested (Carlson & Wiedl, 1992) that students' attempts to verbalize learning difficulties, and receiving elaborated feedback about their efforts, contribute heavily to the value of dynamic assessment. The KMA differs from dynamic assessment procedures because it does not include any of these additional attempts to facilitate learning; students are merely given another opportunity to learn the words from a text passage without any other assistance.

The results of the present research indicate that the opportunity to learn the meanings of some words from the text was most important only in the first study relating the knowledge monitoring procedure to reading comprehension, and seemed to have little effect on studies of college learning or performance estimation. The findings indicated that, with the possible exception of relationships with reading comprehension, use of the word list and vocabulary alone appear to be effective in estimating metacognitive knowledge monitoring, whether the text passage is used or not.
The distinction between explicitly and implicitly defined words was expected to be useful only in those studies in which students read the text passage. The results of those investigations indicated that there were few differences between these types of words. Because neither the use of the text passage nor the distinction between the two types of words seemed to affect the results, it seems prudent to abandon that distinction in future research.

Implications for Training Research and Instruction

The results indicated that use of the text passage did not add much variance to the use of the KMA as an appraisal instrument. It may, nevertheless, be interesting to use the passage in future research to study the applicability of the KMA for research on the training of knowledge monitoring. If the word list and vocabulary test are used as pre-post measures, the text passage could be interspersed to help students learn the meanings of those words about which they had made incorrect knowledge estimates. Different levels of instructional support (Tobias, 1989) could be used to help students learn the meanings of the words they had judged incorrectly.

Use of the text passage makes it possible to implement a training strategy featuring maximal prompting in the form of very active instructional interventions at the beginning and fading those out until the passage alone is presented without any prompts. The interventions could include such procedures as: urging students to provide definitions or synonyms for the words, asking them to rephrase the clauses containing the target words, asking questions about the words, and cueing students that the target words are especially important or that they should pay special attention to them. Of course, research would have to determine whether the suggested interventions actually constitute a hierarchy ranging from maximal to minimal support. It should also be noted that a number of passages, with associated word lists and vocabulary sets, may be needed to develop an effective knowledge monitoring training procedure. Once research has determined the usefulness of the procedures outlined above, they could become an important resource to help teachers at all levels improve the knowledge monitoring of their students.

In addition to the possible usefulness of the instructional interventions described above for training, they could enhance the similarity of the KMA to dynamic types of assessment, and to students’ school learning. Research could then determine whether such interventions improve the knowledge monitoring procedure’s relation-
ship to school learning. It should be noted that, giving students an opportunity for new learning before administering or re-administering, the knowledge estimating procedure is likely to be more complex in mathematics or science than it is for vocabulary. Dynamic assessment in these fields would probably require very active instructional interventions before students can improve their knowledge, because few people can master new material in science or mathematics merely by being asked to read a passage and by twice working on problems in that field, or even by the type of interventions suggested above.

Optimal Indicators of the Latent Knowledge Monitoring Construct

Metacognitive knowledge monitoring is a latent construct inferred from the various scores generated by the procedure. Many of the preceding studies combined the ++ and -- scores to form a measure of knowledge monitoring accuracy. The combined score seemed to have face validity as the most direct and most theoretically interesting index of knowledge monitoring accuracy. Furthermore, by including the -- scores the combined total seemed independent of students' actual knowledge, because the combined estimate included items answered incorrectly. Scores based on the signal detection paradigm were used in Study II, but seemed to add little to the combination of ++ and -- scores used in the other studies. However, the findings of some of the investigations, especially Studies VII and VIII, suggested that differences between groups were obscured when the sub-scores for different categories (++, + -, -+, and -- for words defined explicitly or implicitly) were combined.

Ideally, the optimal knowledge monitoring score should be determined empirically, rather than on the basis of its face validity. The four subscores, or eight if the explicit-implicit distinction is used, generated by the procedure should be submitted to procedures such as the analysis of covariance matrices in order to determine which score(s) are optimal indicators of the latent knowledge monitoring construct. Further research is clearly needed with larger samples (perhaps 200-300 students) than previously employed in order to obtain some stability for the results. The data should then be analyzed with LISREL or comparable procedures in order to identify empirically the optimal score of the latent knowledge monitoring construct.

Knowledge and Estimates of Knowledge

Research has indicated that vocabulary scores are one of the most powerful predictors of school learning (Breland, Jones, & Jenkins,
The knowledge monitoring procedure scores combine both students' estimates of what they know and their actual knowledge. Thus, the + + score is a composite of both word knowledge, determined by the raw score on the vocabulary test, and the students' estimates of that knowledge. Each of the studies described above examined whether the estimates contributed independent variance above that accounted for by students' knowledge. Operationally, this question was analyzed by comparing the variance accounted for by correct estimates (+ + and - - combined) and those representing only the number correct on the vocabulary test (+ + added to - +). Table 7 summarizes these results for each of the studies.

Table 7 indicates that in Studies V, VII, and VIII (four comparisons) knowledge alone, determined by raw score on the vocabulary test, accounted for more variance (ranging from 1-17%) than the estimates. Also, there seemed to be little difference between actual knowledge and estimates in Study III. When knowledge estimates of college students taking introductory psychology classes were related to their Psychology AP scores, the effect size for knowledge alone was 13% (Study VII) and 17% greater (Study VIII) than for knowledge estimates. When relationships between indices of introductory psychology students' in-class performance and KMA scores were analyzed (Study VIII) the effect size for knowledge alone was 6% greater. It is not unusual for vocabulary knowledge, even in an unrelated domain, to be an important predictor of students' grades in college exams, such as the multiple-choice test and the AP examination used in Studies VII and VIII. Vocabulary scores in domains not directly related to the curriculum have been powerful predictors of all types of school learning (Breland, Jones, & Jenkins, 1994; Just & Carpenter, 1987), and findings that they were highly related to how much students learned in a psychology course (determined by either the AP exam or in-class tests) were not surprising. Furthermore, because students had little prior experience with the content of the AP examination they had no basis for estimating their performance on that test. Therefore, in such instances it is reasonable that actual knowledge may be more important in determining students' achievement than estimates of that knowledge.

Knowledge estimates accounted for more variance in seven studies, nine comparisons (ranging in effect size or $R^2$ from 1% to 58%, with a median of 4% more variance), compared to knowledge alone. The largest differences occurred in the study of need for feedback where vocabulary raw score accounted for an insignificant 4% of the variance, and accurate knowledge monitoring estimates accounted
Table 7. Summary Comparing Results for Metacognitive Knowledge Estimates and Actual Knowledge.

Study Results Comparing Metacognitive Estimates (KMA) and Number Correct (Raw scores)

1. KMA accounted for 4% more variance than raw scores.
2. KMA accounted for 5% more variance than raw scores.
3. Correlations similar for KMA and raw scores.
4. Combined GPA differentiated KMA scores, effect size = .07, raw scores NS (effect size = .03). English GPA differentiated KMA scores, effect size = .07, raw scores .06. Humanities GPA differentiated KMA scores, effect size = .09, raw scores .05.
5. Vocational HS Low and High GPA groups differed on KMA, effect size = .14 and on raw scores, effect size = .16. Predicted, postdicted, and actual final exam score = ns for KMA & raw scores.
6. Difference between HS Ss and dropouts greater on KMA, effect size = .13 than on raw scores, effect size = .10.
7. AP data and final grade related to KMA, effect size = .14, and raw scores, effect size = .27.
8. AP data and final grade data related to KMA, effect size = .16, and raw scores, effect size = .33. Class test data related to KMA, effect size = .17, and raw scores, effect size = .23.
9. KMA $r^2$ with Metropolitan score = .58, raw score = .27.
10. Estimates $r^2$ with need for feedback = .62, raw score .04 (ns).
11. Differences between regular, LD, & ADHD Ss greater with KMA than raw scores, effect sizes = .18 compared to .15 for raw scores.

ns = nonsignificant.
* Could not be determined.
for a substantial 62% of the variance! Of course, that finding should be replicated on larger samples. Nevertheless, it seems reasonable that need for feedback should rely more heavily on students’ estimates than on their knowledge.

Another large difference between the contributions of estimated and actual scores occurred in Study IX, one of the math studies. Estimates of the number of problems that could be solved accounted for 31% more variance than the problems actually solved. The findings of Study IX were replicated substantially in Study X; unfortunately a computer malfunction made it impossible to compare the estimated and actual scores in that investigation. Although the math studies clearly need replication, the findings suggest that knowledge estimates may be more powerful predictors of success in that domain than in vocabulary.

One possible reason for the substantial effects in mathematics compared to vocabulary may deal with domain similarity. That is, knowledge estimates in math were made from content that was highly similar to the types of problems encountered during math instruction. As indicated above in the discussion of the performance expectation studies, the vocabulary words used in the research were not similar to the domains in which instruction occurred, or to other types of external criteria, perhaps leading to somewhat weaker effects. That interpretation is supported by findings from several of the investigations. In Study I, relating the declarative word knowledge and estimates of that knowledge to reading comprehension, the highest relationships were found for KMA scores after students had read the text passage in which the vocabulary words were defined. That procedure was obviously very similar to the task students face in reading comprehension tests. Furthermore, in Study III social science and science had the lowest relationships with KMA scores, and the effects for social science and science were insignificant in Study IV. Because the KMA materials were developed to be quite general, they were probably dissimilar to what students learned in these more technical areas. These results suggest that the KMA has stronger effects within a domain, rather than across domains. Schraw, Dunkle, Bendixen, and Roedel (1995) found that knowledge monitoring had both domain specific and domain general attributes. Further research is needed to clarify the domain specific and/or domain general characteristics of the KMA.

Another possibility accounting for the more positive results in the studies involving mathematics deals with the perceived difficulty of the subject. Everson, Tobias, Hartman, and Gourgey (1993) found
that students perceive mathematics to be the second most difficult subject, right after science. Conceivably, as suggested below, students' estimates of their knowledge in more difficult domains are less automatic and involve more reflection about their prior experiences than in simpler areas. Students' confidence and/or their anxiety about these fields may also affect their estimates. Further research is clearly needed using materials drawn from mathematics, science, and other fields to study both this question and the issue of domain generality-specificity.

Difficulty and Knowledge Monitoring Procedure

Little information about the difficulty of the various vocabulary and mathematical materials was available prior to their use in any of the studies. This may well have contributed to some of the variable results. It seems reasonable that estimates of knowledge based on students' thoughtful consideration of what they know and do not know would be more substantially related to other variables than estimates made more or less automatically. Rapid answers made with little reflection are most likely when students respond to materials that are very easy for them. Wrong estimates for such relatively automatic responses probably indicate careless errors, rather than failures of well-considered estimates. More difficult materials may also evoke nonreflective responses, because students may feel that they neither know nor care about what the correct answers to such questions are. Items of moderate difficulty, about which students may have partial knowledge that can be extended by exerting some effort, would appear to be most likely to elicit well-considered responses reflective of students' metacognitive knowledge monitoring ability.

Item difficulty is also of importance in considering the different KMA scores. Of the four scores generated by the procedure, the greatest number of responses fell into the ++ category in the studies described above. It may be assumed that more difficult items would yield more -- and + + responses, increasing their reliability and the likelihood that they could contribute more variance to the discrimination between accurate knowledge monitors and their less accurate peers. Furthermore, having more items in the -- category will reduce the similarity between estimates and number correct for two reasons: First, such response represents accurate estimates but no knowledge about the item, and second, more -- items leaves a smaller percentage of ++ items.

In future research these expectations about the effects of varying item difficulty levels should be tested by using items with a previ-
ously determined range of known difficulty. It could be hypothesized that the most useful metacognitive knowledge estimates are likely to be generated from materials of moderate difficulty, and that more difficult items will increase the distinction between KMA accuracy and number correct on any of these procedures.

Relationship to Metamemory Research

The procedure described in this chapter is similar to metamemory research on the feeling of knowing (FOK) and judgment of learning (JOL). FOK judgments "occur during or after acquisition and are judgments about whether a given currently non-recallable item is known and/or will be remembered on a subsequent retention test....Judgments of learning (JOL) occur during or after acquisition and are predictors about future test performance on currently recallable items" (Nelson & Narens, 1990, p. 130). In terms of that definition, students' judgments on both the word list and math problems in the preceding research were similar to JOLs.

FOK research was originated by Hart (1965) who asked general information questions of students who, after failing to recall an item, had to make a judgment regarding their FOK about that item. Finally, they were asked to select an answer from a set of distractors. The procedure has been extended to asking students to guess if they could recall words learned in a paired associate task (Hart, 1967; Ryan, Petty, & Wentzlafl, 1982). Nelson, Gerler, and Narens (1984) also extended the FOK research to students' ability to relearn, and to perceptual identification tasks. Reder and Ritter (1992) investigated whether students opted either to retrieve or calculate mathematical problems, and the latency and accuracy of these processes. A review of FOK research indicated that "a large number of studies confirmed that (students)....unable to retrieve a solicited item from memory can estimate with above chance success whether they will be able to recall it in the future, produce it in response to clues, or identify it among distractors....The standard finding is that the predictive validity of FOK judgments is above chance, though far from perfect" (Koriat, 1993, p. 609-610).

The FOK and JOL paradigms differ from the present research in a number of ways. First, the FOK judgments are typically required after a recall failure, rather than after every stimulus presentation. Second, in FOK or JOL research no attempts are usually made to enable students to learn and/or correct their knowledge of the stimuli, as they were in some of the present research. Third, the purposes of
the metamemory research are to clarify the mechanisms accounting for FOK and JOL, rather than to use the scores as a measure of metacognitive knowledge monitoring to be related to different variables of importance in students' school learning.

Suggestions for Further Research

A number of recommendations for further research have been made earlier; additional suggestions that do not pertain directly to the previous discussion are made here. The positive findings relating knowledge monitoring to need for feedback suggest that studies of similar variables relating the procedure to processes of importance in school learning may be fruitful. For example, forgetting what has been learned in school may be related to knowledge monitoring. It could be inferred that students with good knowledge monitoring abilities, by having a clear sense of what they know and do not know, may be able to retrieve more prior learning than those who have a less secure grasp of what they know and do not know and, hence, may have greater difficulty retrieving prior learning. A pilot study of the knowledge monitoring-forgetting relationship provided substantial support for that reasoning, and will soon be followed up.

The relationship between knowledge monitoring and the effect of distractibility is another fruitful area for investigation. Even though there is a great deal of anecdotal evidence that students are readily distracted from their studies, it has been surprisingly difficult to divert students in investigations specifically designed for that purpose (Slater, 1968; Tobias, 1973). Although some of that variability may be attributable to motivational phenomena (i.e., the interest level of both the primary and distracting materials seems to be important in determining whether students are successfully diverted from their studying; Tobias, 1973), students' knowledge monitoring abilities may also help to determine whether students are distracted. Students with an accurate grasp of their knowledge should find distractions less disruptive from their work than those with a hazier notion of what they know and do not know.

Research should also be conducted relating knowledge monitoring to depth of processing (Craik & Lockhart, 1972). Students should be able to distinguish between the known and unknown more accurately if the learning was processed at a deep, rather than shallow level. Deeper processing should enhance students' knowledge monitoring ability, and it could be predicted that students will make more accurate distinctions between the known and unknown on material
they are induced to process deeply, either by experimental manipulations or instructions, rather than at a shallow level.

Learning in complex domains, such as science, engineering, or making diagnoses in medicine or other fields, often requires that students bring substantial amounts of prior learning to bear in order to understand and acquire new knowledge, and/or solve problems. Some prior learning may be recalled imperfectly, or may never have been completely mastered during initial acquisition. Students who can accurately distinguish between what they know and do not know should be at an advantage while working in such domains, because they are likely to review and try to relearn imperfectly mastered materials needed for particular tasks more readily than students who are less accurate in making such differentiations.

REFERENCES


to students with low mathematics ability. Teaching and Teacher
Education, 8, 109–121.

assessment: The application of a specific model. Learning and Individual Differences, 4, 153–166.


Metacognition refers to thinking about thinking, or more generally, to using higher-level knowledge and strategies to regulate lower-level performance. Previous research suggests that metacognition is an important part of learning among adults (Baker, 1989; Garner & Alexander, 1989; Pressley & Ghatala, 1990) and children (Alexander, Carr, & Schwanenflugel, 1995; Borkowski & Muthukrishna, 1992). Metacognition contributes to learning in several ways, but especially by helping learners to use their attentional resources more efficiently, to process information at a deeper level, and to monitor their performance more accurately.

Notwithstanding its importance, there is considerable debate regarding how to measure metacognition. At the heart of the problem is the elusive nature of metacognitive knowledge itself. Most theorists assume metacognitive knowledge is highly abstract and cuts across domain-specific boundaries (Brown, 1987; Flavell, 1987; Paris & Byrnes, 1989; Schraw, Dunkle, Bendixen, & Roedel, 1995; Schraw & Moshman, 1995). In contrast, most declarative and procedural knowledge in memory is welded to a specific domain, and can be stated as a declarative fact or demonstrated through a procedure. As a result, declarative and procedural knowledge are much easier to identify, manipulate, and measure than metacognitive knowledge. Added to
this is the fact that metacognitive knowledge is acquired gradually over long periods of time, emerges relatively late in development, and often is difficult to explicate even when an individual demonstrates a high degree of metacognitive competence (Brown, 1987; Garner, 1994; Weinert & Kluwe, 1987).

Another problem is that metacognitive processes such as planning and evaluation are difficult to measure directly.

For this reason, researchers have relied on a variety of indirect measures such as verbal reports, think-alouds, self-report inventories, and subjective measures of performance accuracy. One consequence of the unobservable nature of metacognitive knowledge and regulation is that researchers have focused their attention on several specific aspects of metacognition that are easier to measure than others, especially various forms of monitoring. Most studies have focused on memory monitoring (Cavanaugh & Perlmutter, 1982; Johnson, Hastroud, & Lindsay, 1994; Lovelace, 1984; Koriat, 1993; Schneider & Pressley, 1989), comprehension monitoring (Glenberg & Epstein, 1985; Leonesio & Nelson, 1990; Weaver, 1990), or performance monitoring (Glenberg, Sanocki, Epstein, & Morris, 1987; Pressley & Ghatala, 1990).

This chapter addresses problems related to the measurement of metacognition in greater detail. We believe that some of the more imposing obstacles can be addressed successfully via computer-based testing procedures, but especially those pertaining to the assessment of metacognitive control processes. We will argue that computer-based testing provides opportunities for researchers to measure control processes with much greater precision than with noncomputerized methodologies. Computer-based testing enables us to do so in an unobtrusive, reliable manner that is less apt to be confounded by preexperimental knowledge and ability.

The remainder of this chapter is divided into six sections. The first of these provides a brief overview of previous research and presents a multilevel model of metacognition that distinguishes between two major components, including knowledge about and regulation of cognitive processes and knowledge. We further distinguish between two subcomponents of metacognitive regulation, including metacognitive control and monitoring. Control processes are used to select performance goals and guide ongoing cognitive activities. Monitoring processes are used to evaluate the present success of one’s performance and the degree to which one has met one’s long-term performance goals. We assume that control and monitoring are reciprocally linked in a manner that facilitates self-regulation during performance.
The overview is followed by a section that outlines some of the methodological shortcomings of previous research. These include issues pertaining to the reliability and construct validity of dependent variables used in these studies. Of greater importance, this section considers how dependence on a limited repertoire of methodological strategies has precluded inquiry along two important lines. The first concerns the investigation of metacognitive control. We believe that few studies have investigated control processes at all, and that none have done so directly. The second line of inquiry concerns the relationship between control and monitoring processes. Current conceptualizations of metacognition make a number of assumptions about this relationship that have not been tested empirically.

The next section provides a review of recent developments in computer-based testing that offer great promise for the measurement of metacognition. These include the contribution of item-response theory to the rapidly growing field of computerized adaptive testing (i.e., tests in which a computer-controlled algorithm selects test items from a multilevel, calibrated item pool) and self-adapted testing (i.e., tests in which examinees select item of a designated difficulty level from a multilevel, calibrated item pool).

We consider ways that self-adapted testing (SAT) can be applied to the measurement of metacognition in the next section. This includes some of the psychometric advantages of SAT as well as a description of on-line measures of cognitive and metacognitive behavior that can be used to test the model of metacognition proposed later in this chapter. Specifically, we address how SAT can be used to assess metacognitive control in a variety of ways, including a measure of how accurately individuals select test items, as well as selection times, item response times, and across-test item selection strategies.

The final section outlines an agenda for future research using SAT. One important goal of this research is to link the kinds of data collected in previous studies with the kind of on-line measures available in SAT. Among other things, this would enable researchers to compare the reliability of subjective paper-and-pencil judgments made before, during, or after testing to objective measures collected during SAT. Ideally, one would hope for a strong correspondence between the two; however, one possibility is that pre- and post-test subjective judgments do not correspond closely to actual on-line item selection strategies. Another goal is that researchers investigate in detail the relationship between control and monitoring. One would expect these processes to be linked reciprocally, even though there is no direct empirical evidence to support this assumption. Establishing
such a relationship would suggest that control and monitoring processes are part of a larger regulatory system. In contrast, finding that the two processes are not related strongly would suggest that each is governed by a separate reservoir of knowledge.

The final section of the paper summarizes our main points and offers some general conclusions. Chief among these is the claim that researchers may benefit by incorporating recent innovations from the computer-based testing community, and by using SAT to bridge the gap between existing metacognitive theory and empirical studies that do not adequately address questions raised by this theory.

**COGNITIVE AND METACOGNITIVE PROCESSES**

Individuals rely on both cognitive and metacognitive skills when learning (Garner & Alexander, 1989; Pressley, Borkowski, & Schneider, 1987). Cognitive skills are those that help a person perform a task; metacognitive skills are those that help a person regulate and monitor task performance (Artzt & Armour-Thomas, 1992; Schraw, 1994; Slife & Weaver, 1992). Metacognition is thought to include two main components (Baker & Brown, 1984; Brown, 1987; Jacobs & Paris, 1987). The first, knowledge of cognition, refers to what individuals know about their own cognition or about cognition in general. It usually includes three different kinds of metacognitive awareness: declarative, procedural, and conditional knowledge (Brown, 1987; Jacobs & Paris, 1987; Schraw & Moshman, 1995). Declarative knowledge refers to knowing “about” things. Procedural knowledge refers to knowing “how” to do things. Conditional knowledge refers to knowing the “why” and “when” aspects of cognition. The second, regulation of cognition, refers to metacognitive activities that help control and monitor one’s learning. Although a number of regulatory skills have been described in the literature (Jacobs & Paris, 1987; Kluwe, 1987), two that appear to be essential are control and monitoring processes.

A growing number of studies have been conducted over the past decade investigating these components. Those focusing on the knowledge of cognition component typically employed either think-aloud (Swanson, 1990) or self-report measures (Dixon, Hultsch, & Hertzog, 1988; Schraw & Dennison, 1994). Those focusing on the regulation of cognition component, but especially the monitoring subcomponent, typically employed some form of priming task, or asked individuals to make subjective judgments of confidence, ease of comprehension, or overall learning prior to or subsequent to completing a test.

Unfortunately, because many of these studies used widely different materials, data collection procedures, and criterion measures, results are mixed and often difficult to compare. In lieu of a comprehensive review of these diverse findings, we turn briefly to a summary of recent research investigating the control and monitoring subcomponents. We do so for two reasons. One is to provide a more detailed definition of each construct. A second is to delineate the strengths and weaknesses of recent empirical research.

Research on Control

Metacognitive control refers to regulatory processes that occur prior to or during a learning activity that direct the course of cognitive activities. These processes include but are not restricted to planning, allocating resources, selecting strategies, and setting specific performance goals. Control processes typically are assumed to guide cognitive activities in a top-down manner (Nelson & Narens, 1990, 1994). Most theorists also assume that control processes are intentional, nonautomated, and partially stateable (Bjorklund & Harnishfeger, 1990; Pressley, Harris, & Marks, 1992).

Many studies have investigated the effect of strategy instruction on metacognitive control (see Garner, 1987; Pressley et al., 1987; Pressley, 1995, for reviews). These studies invariably indicate that strategy instruction increases metacognitive control in two ways: through better use of limited cognitive resources and more elaborative processing (Willoughby, Wood, & Khan, 1994; Wood, Pressley, & Winne, 1990). However, few of these studies have shown attempts to assess the accuracy of enhanced control processes, the degree to which learners have metacognitive awareness about enhanced control, and the extent to which enhanced control is related to monitoring accuracy.

Several studies have investigated the relationship between control and monitoring more directly. Pressley and colleagues (see Pressley & Ghatala, 1990, for a review) found that experimental manipulations that improved performance (presumably by enhancing metacognitive control) did not lead to more accurate monitoring among college students. In contrast, Maki and colleagues (Maki & Serra, 1992; Maki, Foley, Kajer, Thompson, & Willert, 1990) found that experimental manipulations that neces-
situated deeper information processing (e.g., asking readers to generate missing text information) led to more accurate monitoring.

Other studies have used estimates of future performance on a specific task as a measure of metacognitive control. Schraw (1994) asked college students to estimate their ability to monitor accurately their reading comprehension. Control predictions were correlated positively (i.e., $p < .01$) with test performance and post-test estimates of monitoring accuracy. Levels of self-assessed monitoring ability also were related to item-by-item and end-of-test monitoring accuracy. Those who rated themselves as normatively accurate monitors tended to be more accurate and to improve more than poor monitors as a function of self-generated feedback. These findings suggested that older learners possess knowledge about metacognitive processes and use this knowledge strategically to control their performance and monitoring.

A follow-up study by Schraw (1995) examined performance control judgments (i.e., pretest estimates of one’s ability to perform well in a specific domain) across a variety of content domains and test formats. Results indicated that control judgments were correlated positively among domains even when test performance was controlled statistically. This suggested that metacognitive control may be a domain-general rather than domain-specific phenomenon. However, control judgments across different types of tests (i.e., recognition of facts versus recall of inferential relationships) were unrelated. This suggested that control judgments may be dependent on the specific cognitive processes required of a particular test format (see Pressley & Ghatala, 1990, and Schwartz & Metcalfe, 1994, for a further discussion).

Research on Monitoring

Metacognitive monitoring refers to processes that occur during or after a learning activity that provide information about the effectiveness of those activities. These processes are used to evaluate the present success of one’s performance and the degree to which one has met one’s long-term performance goals. Monitoring is important because it provides self-generated feedback to the control system. Without accurate monitoring, efficient control of one’s performance may be impossible. Most theorists assume that monitoring is a data-driven process; that is, monitoring accuracy may be a function of domain familiarity, automaticity, and task difficulty (Koriat, 1993; Nelson & Narens, 1990, 1994).
Monitoring studies typically require individuals to make subjective judgments of learning or test performance during or after an initial study phase. Judgments are made for each test item using a 5- or 7-point Likert scale, although some researchers have used other techniques such as a continuous, bipolar scale adapted from the multidimensional literature (see Schraw, Potenza, & Nebelsick-Gullet, 1993, for a further description). The main purpose of these studies is to determine the degree to which individuals accurately assess their learning and performance.

Four types of judgments have been used in the adult monitoring literature, including ease of learning (i.e., judgments of encoding difficulty), judgments of learning (i.e., the degree to which information was learned during the study phase), feeling of knowing (i.e., the degree to which one has access to previously learned information in memory), and performance judgments (i.e., assessments of performance accuracy). These four types of judgments have been used by researchers to operationalize metacognitive processes involved in the acquisition, retention, and retrieval of information (Nelson & Narens, 1994).

Monitoring studies differ widely with respect to the type of criterion measure used to assess monitoring ability. Many studies use some form of correlation, although a number of studies report other measures such as bias (Schraw & Roedel, 1994), accuracy (Tobias, 1996), discrimination (Lundeberg, Fox, & Puncochar, 1994), or a multicomponent measure based on bias, correlation, and discrimination (Yates, 1990). Currently, there is widespread disagreement about the relative effectiveness of these measures (Keren, 1991; Liberman & Tversky, 1993; Nelson, 1984; Schraw, 1995). One point of agreement is that different criterion measures affect both observed results and how researchers interpret these results.

These studies generally suggest that adults monitor their learning and performance with a moderate degree of success, although results vary from study to study. Surprisingly, monitoring proficiency does not appear to be related strongly to relevant domain knowledge (Glenberg & Epstein, 1987; Morris, 1990; Schraw et al., 1995) or academic achievement (Pressley & Ghatala, 1988, 1990). These conclusions have been supported in the children's monitoring literature as well, although there is considerable debate regarding whether children monitor as accurately as adults (Alexander, Carr, & Schwanenflugel, 1995; Butterfield, Nelson, & Peck, 1988).

Situational constraints also affect estimates of monitoring proficiency. One constraint is the point in the learning-test sequence in which monitoring judgments are made. A number of studies indicate
that calibration of comprehension (i.e., the correlation between pretest judgments and actual test performance) is often quite poor, with most studies reporting correlations in the .00 to .25 range (Glenberg et al., 1987; Pressley & Ghatala, 1990). In contrast, calibration of performance (i.e., the correlation between posttest judgments and actual test performance) appears to be much better in both children and adults, often ranging from .30 to .50 (Glenberg et al., 1987; Maki & Serra, 1992; Maki et al., 1990; Pressley & Ghatala, 1990).

A second constraint is that specific testing conditions affect monitoring proficiency. For example, calibration of comprehension can be improved under the following circumstances: (a) when adjunct questions similar to post-test questions are provided during study (Pressley, Snyder, Levin, Murray, & Ghatala, 1987), (b) when periodic feedback is provided to test takers (Ghatala, Levin, Foorman, & Pressley, 1989), (c) when expert knowledge about the to-be-learned material is minimized (Glenberg & Epstein, 1987), and (d) when test takers generated missing text information (Maki et al., 1990). Surprisingly, calibration of comprehension does not appear to improve when learners were specifically requested to monitor their comprehension or when they are given the opportunity to re-study the to-be-learned materials (Ghatala et al., 1989), or when they were given practice questions prior to study (Maki & Serra, 1992).

Like calibration of comprehension, calibration of performance improved under a number of testing conditions, especially when adjunct questions were provided during the study phase (Pressley et al., 1988), when test takers received external incentives to improve monitoring accuracy (Schraw et al., 1994), and when test takers received recall rather than recognition tests (Pressley, Ghatala, Woloshyn, & Pirie, 1990). Calibration of performance also was related to level of test performance (Schraw & Roedel, 1994). Individuals monitored with less bias when judging their performance on easy rather than more difficult items.

A third general constraint is that monitoring proficiency improves with feedback, incentives, practice, and training. Stock, Kulhavy, Pridemore, and Krug (1992) found that experimenter-provided feedback increased the accuracy of confidence judgments. Schraw (1994) reported that pre-experimental estimates of monitoring proficiency were related to both local (i.e., the accuracy of item-specific performance judgments made during testing) and global (i.e., judgments of overall performance made after testing) monitoring accuracy. The accuracy of local monitoring was correlated positively to the accuracy of global monitoring. In addition, the change in
monitoring accuracy between local and global monitoring improved significantly among good monitors, but did not improve among poor monitors.

Monitoring proficiency also improves when individuals are given incentives to monitor their performance more accurately. Schraw et al. (1993) found that additional course credit for normatively high monitoring accuracy led to more accurate monitoring, whereas additional credit for normatively high test performance had no effect on monitoring accuracy. In addition, incentives to monitor more accurately improved test performance even though incentives to perform better did not.

Monitoring training also improves performance. Delclos and Harrington (1991) examined fifth and sixth grader’s ability to solve computer problems after assignment to one of three conditions. The first group received specific problem-solving training, the second received problem-solving plus self-monitoring training and practice, and the third received no training. The monitored problem-solving group solved more of the difficult problems than either of the remaining groups and took less time to do so. The group receiving problem-solving and monitoring training also solved complex problems faster than the control group.

Summary

The control and monitoring research summarized above leads to a number of conclusions. Regarding control, most adults achieve some degree of metacognitive control by using helpful learning strategies. Second, many adults possess some explicit metacognitive knowledge about their ability to control performance. Third, metacognitive control in one domain tends to be related to control in another domain, even when performance is taken into consideration. Fourth, metacognitive control appears to be superior in adults (Alexander et al., 1995).

Regarding monitoring, adults monitor their performance with a moderate degree of accuracy. Monitoring improves as tests become easier and more factual. Second, monitoring proficiency appears to be independent of intellectual ability (Alexander et al., 1995; Swanson, 1990) and academic achievement (Pressley & Ghatala, 1988). Third, monitoring proficiency may be independent or even negatively related to domain knowledge (Glenberg & Epstein, 1987), independent of ease of comprehension judgments (Leonesio & Nelson, 1990), but correlated with other types of metacognitive knowledge (Schraw,
PROBLEMS WITH CURRENT MEASUREMENT APPROACHES

It could be argued that the gap between metacognitive theory and empirical research is as great as any other area of psychological inquiry. These are several specific reasons for this state of affairs, many of them being methodological in nature (Kruglanski, 1989). This section divides these problems into three interrelated categories that are ranked ordered from our vantage point in order of importance. The three categories include task, test, and person constraints on the measurement of control and monitoring.

Task Constraints

Task constraints refer to characteristics of the experimental task that impede measurement of either control or monitoring processes. The most serious obstacle is that researchers cannot manipulate either control or monitoring processes directly, but must be content to manipulate the task environment in which control and monitoring are performed. This means that researchers must make inferences about complex metacognitive processes on the basis of indirect measures. Although this is certainly not a new problem to psychologists, it is a serious one.

Operationalizing metacognitive control has been an especially virulent problem. Presumably, the best way to study control processes would be to allow the examinee to exercise a great deal of strategic control over his or her performance. Previous studies have attempted to do so by providing specific task information, learning goals, opportunity to study, strategies for learning, or conditions under which learning is facilitated. In essence, these studies examined whether a variety of experimental factors affected metacognitive control. However, none of these studies allowed examinees to demonstrate overtly in a directly observable manner how they attempted to control their test-taking behavior. One way to do so would be through the use of on-line verbal protocols in which individuals describe their cognitive processes (Ericsson & Simon, 1984; Pressley & Afflerbach, 1995). However, although an important research tool, verbal reports are intrusive, resource consuming, and assume that individuals have explicit access to metacognitive processes.

An alternative would be to study the way examinees make strategic choices throughout a test. In self-adapted testing, for ex-
ample, individuals choose test items of a designated difficulty level from a multilevel, calibrated item pool. This may enable researchers to examine several aspects of metacognitive control in an explicit, yet unobtrusive manner. One aspect is the goodness of fit (i.e., calibration accuracy) between self-selected items and observed performance. Another aspect is whether examinees show evidence of improved accuracy over the course of the entire test.

A somewhat different task constraint is introduced when researchers ask examinees to make subjective judgments of learning and performance while simultaneously performing complex tasks. Researchers invariably assume that such ratings have little effect on performance, although oddly, there are no empirical studies we know of that have investigated this assumption. Of greater importance, researchers also assume that the demands of taking a test have little impact on the accuracy of subjective ratings. This assumption clearly is untenable in that confidence judgments become increasingly more biased as a function of test difficulty (Schraw & Roedel, 1994; Schwartz & Metcalfe, 1994). Although researchers have attempted to compensate for such problems via the judicious use of statistical analyses (cf. Nelson, 1984), no amount of statistical tinkering can eliminate these problems entirely (cf. Funder, 1987; Keren, 1991; Liberman & Tversky, 1993; Schraw, 1995).

Test Constraints

Test constraints refer to characteristics of the test itself, rather than the test environment, that impede measurement of either control or monitoring processes. A recent review by Schwartz and Metcalfe (1994) addressed four test-related problems that we summarize here. One source of variation among examinees, and presumably an important source of measurement error, pertains to the type of test being given. Recall tests often are assumed to be more cognitively demanding than recognition tests. Most empirical studies echo this difference by revealing higher correlations between performance and confidence (or accuracy) judgments on recall tests. One reason for higher correlations is less restriction of the range of scores on recall tests when compared to recognition tests. Because recall tests are more difficult, their scores will vary across a wider range of possible values. In contrast, easier recognition tests restrict the observed range of a correlation due to homogeneous performance or ceiling effects.

Another inadvertent problem of recognition tests is that examinees are influenced by the availability of information included in the
test item. Because test answers are provided explicitly in a recognition test, but must be generated in a recall test, examinees are significantly more confident when monitoring recognition tests, but more accurate when monitoring recall tests (Ghatala, Levin, Foorman, & Pressley, 1989).

A second major source of measurement error is the length of a test, or if it is a recognition test, the number of alternatives from which one may choose for each item. It is well established that a test’s reliability is directly related to its length, with longer tests, and recognition tests with more alternatives, being more reliable (Crocker & Algina, 1986). Unfortunately, many early studies of monitoring used multiple tests with one or two items per test, rather than the preferable one test with a large number of items. To illustrate, Glenberg et al. (1987) reported no statistically significant relationship between pretest judgments of learning and subsequent performance. This group of experiments required individuals to answer one main idea question per test for a large number of tests. Replicating this study, having first increased the length of each test, Weaver (1990) found that the observed value of $r$ increased monotonically as a function of test length, until it reached an asymptotic value of $r = .60$. Thus, Glenberg et al. (1987) failed to identify a significant relationship between judgments of learning and test performance due to unreliable test scores.

A third source of error is test difficulty. Monitoring accuracy declines as a test becomes more difficult, even when test performance is controlled statistically (Schraw & Roedel, 1994; Schraw, Dunkle, Bendixen, & Roedel, 1995). In addition, overconfidence is more common than underconfidence and more likely to occur when a test is difficult (Cutler & Wolfe, 1989; Newman, 1984). These patterns have been observed on a variety of tasks including probability judgments (Fischhoff, 1988), reading comprehension (Glenberg et al., 1987), recalling emotions (Thomas & Diener, 1990), and social judgments (Dunning, Griffin, Milojkovic, & Ross, 1990).

There are at least two reasons a difficult test may interfere with control and monitoring processes. One is that individuals lack sufficient background knowledge to answer the test question. It is well known that individuals resort to a number of helpful, but fallible, heuristics under these circumstances that bias their judgments (Fischhoff, 1988; Tversky & Kahneman, 1973). A second reason is that information in memory is inaccessible during testing (i.e., available in memory, but presently unretrievable). Partial or total inaccessibility may lead to severe judgment bias due not only to poor monitoring, but fallible retrieval processes as well (Koriat, 1993, 1994).
A fourth source of error is knowledge about the test. Test-relevant knowledge may affect control and monitoring in several ways—namely, by enabling examinees to identify test-relevant information more efficiently, process information at a test-appropriate level (McDaniel & Einstein, 1989), and utilize self-generated feedback (Glenberg et al., 1987). In general, as knowledge of the test increases, performance and the reliability of tests improve as well (Schwartz & Metcalfe, 1994). Research by Metcalfe (1993) also found that administering a test that was not expected reduced the correlation between performance judgments and actual performance dramatically.

**Person Constraints**

There are a number of ways that prior knowledge might affect control and monitoring processes negatively, and thereby reduce the reliability of measurements (Baker, 1989; Garner & Alexander, 1989). Insufficient knowledge may preclude the use of helpful learning and test strategies and lead to lower performance. Lower performance may, in turn, lead to a restriction in the range of observed test scores. Low domain knowledge also makes a test more difficult, which has several deleterious effects on monitoring already described above.

It is possible that prior knowledge interacts with many of the constraints described above in complex ways. For example, low prior knowledge presumably affects the degree to which individuals learn information during a pretest study session. Poorer learning leads to a greater amount of inaccessible information and a more difficult test. Low prior knowledge in a domain also may restrict deeper information processing that could affect performance on some test questions, but not others.

It is important to note, however, that increasing prior knowledge per se does not seem to improve monitoring (Nelson & Narens, 1990; Pressley et al., 1990), unless the inclusion of prior knowledge provides an opportunity for self-generated feedback or additional knowledge about the test itself (Glenberg et al., 1987). For example, research by Morris (1990) found that although knowledge was related positively to performance, it was not related to monitoring accuracy. Schraw (Schraw & Roedel, 1994; Schraw et al., 1995) extended these findings across multiple domains, arguing that individuals possess a domain-general (i.e., knowledge-independent) monitoring skill that is independent of domain knowledge. Glenberg and Epstein (1987) also reported that higher levels of expert knowledge actually interfered with accurate monitoring.
Summary

Empirical studies of control and monitoring lag behind metacognitive theory. One important reason is that each of these processes is difficult to operationalize experimentally and to manipulate directly. Researchers have relied on several limited measurement paradigms, including error detection (see Baker & Cerro, this volume) and subjective calibration judgments. Both of these methodologies are fraught with measurement problems related to the nature of the task itself, to factors including the type and difficulty of the test, and to characteristics of the examinee.

In subsequent sections of this chapter we argue that self-adapted testing allows researchers to eliminate many of these problems, and thereby increase the construct validity of tests (Rocklin, O’Donnell, & Holst, 1995), by (a) controlling for test and item difficulty using a calibrated pool of independent test items, (b) reducing measurement error attributable to characteristics of the examinees such as ability and prior knowledge, (c) utilizing unobtrusive measures that do not compete for the examinees’ limited resources, and most importantly, (d) allowing the test taker to exercise a much greater degree of control during the testing process. We turn now to a brief overview of computer-based testing and two recent developments: computerized adaptive and self-adapted testing.

NEW DEVELOPMENTS IN EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Item Response Theory

During the past few decades, Item Response Theory (IRT), has emerged as the psychometric model used by an increasing number of testing programs in education and psychology. For large-scale achievement and proficiency tests in particular, IRT has largely supplanted classical test theory as the basis for test development, scoring, and equating. The central concept of IRT is the item characteristic curve (ICC), which specifies the relationship between the level of an examinee’s proficiency (i.e., estimated ability) and the probability that he or she passes the item.

The most commonly used IRT models assume that there is a monotonic relationship between examinee proficiency and the probability of passing an item. In addition, it is assumed that the set of test items under consideration is unidimensional (i.e., measures a single, unobservable construct). It has been typically found, however, that
the IRT model will adequately fit the test data if there is one sufficiently dominant factor underlying the items. A detailed explanation of IRT is beyond the scope of this chapter; the interested reader is referred to Hambleton, Swaminathan, and Rogers (1991) for a good overview of basic IRT concepts.

Two principles of IRT are particularly relevant to the present discussion. The first is a key property of IRT, termed invariance, which states that an examinee's proficiency is independent of the characteristics of the items that are administered. Consider the case in which there is an available pool of 400 test items, and that it is larger than would be administered to a given examinee (e.g., 100 items) during a testing session. Regardless of which 100 items were administered from the pool, the examinee's expected proficiency estimate would be invariant. Invariance holds because IRT-based proficiency estimates take into account both (a) characteristics (primarily difficulty) of the items that were administered and (b) the examinee's performance on those items. An important implication of invariance is that two examinees can receive completely different tests, drawn from the same item pool, yet their proficiency estimates can be compared. Any differences in difficulty of the two tests are taken into account by the IRT estimation procedure.

It should be noted that invariance is not a feature of the classical test theory measurement model, in which proficiency estimation is based solely on test performance (i.e., number of items passed). If two examinees take two tests that differ in difficulty, then the difference between the examinees' proficiency levels is confounded with the difference between the difficulty of the tests.

A second principle of IRT that is of particular relevance to the study of examinee monitoring and control is that the difficulty parameters of the ICCs, which indicate the relative difficulties of the items, are placed on the same scale as examinee proficiency. This joint scaling is depicted in Figure 1, which indicates that Item 1 is the least difficult item, followed by Item 2, and so on through Item 5. Moreover, examinee A is the least proficient of the three examinees, and Examinee C the most proficient.

Measuring difficulty and proficiency on the same scale allows one to assess the degree of match between the difficulty of an item and an examinee's proficiency. Why is this joint scaling important? The closer the match between an item's difficulty and an examinee's proficiency, the more informative is the examinee's response to that item in estimating his/her proficiency. Hence, more difficult items are most informative for more proficient examinees, whereas less
difficult items are most informative for less proficient examinees. In Figure 1, the most informative items for Examinees A, B, and C are Items 2, 3, and 5, respectively.

Figure 1. The joint scaling of item difficulty and examinee proficiency in IRT.

---

**Computer-Based Testing**

With the introduction and rapid proliferation of microcomputers came an increased use of computers to administer tests. There are a number of advantages realized with computer-based testing that may make it attractive to examinees, including the capability for on-demand testing, as well as immediate test scoring and reporting of results. From a researcher’s standpoint, however, computer-based testing provides two additional advantages. First, it allows a much greater degree of control over the test administration. Such control may include (a) the order in which items are considered and answered, (b) how long each item is presented, and (c) whether or not examinees are allowed to review, and possibly change, their answers to items. Second, it allows the researcher to unobtrusively collect a great deal of information about the test session, such as how long an examinee took to respond to each item or whether or not the examinee changed his/her answers to any test items. Because of these advantages, a computer-based test provides a unique opportunity for re-
searchers to study examinee test-taking behavior. With paper-and-pencil tests, such advantages are unavailable.

**Computerized Adaptive Testing.** Computerized adaptive testing (CAT) combines the psychometric advantages of IRT with the computing power of current microcomputers. In a CAT, a computer algorithm is used to match the difficulty of the items administered to the estimated proficiency of each examinee. At each step in a CAT, the next item to be administered is a function of the examinee's responses to previously administered items. Using a CAT, examinee ability is estimated more efficiently than with a conventional test because typically fewer items are required to attain the same degree of measurement precision. It has typically been found that a CAT requires about half as many items to estimate an examinee's proficiency with the same degree of precision as a paper-and-pencil test.

Note that both of the IRT principles discussed earlier are essential to a CAT. Because item difficulty and examinee proficiency are on the same scale, items having difficulties matching an examinee's current proficiency estimate can readily be identified and administered. And, because examinees receive unique tests, the invariance property allows their proficiency estimates to be compared.

**Self-Adapted Testing.** Although CAT is by far the most popular application of IRT in computer-based testing, other applications have been studied. One of these is self-adapted testing (Rocklin & O'Donnell, 1987). A SAT is similar to a CAT with one important exception. In a self-adapted test, the examinee is allowed to choose the difficulty level of each test item administered, whereas in a CAT a computer algorithm chooses each item to be administered based on the examinee's performance on items administered earlier in the testing session.

In a SAT, an examinee chooses the difficulty level of each item administered from an item pool has been divided into several (typically 5–8) ordered difficulty levels, or strata, based on the IRT difficulty parameters of the items. This relationship among difficulty levels is illustrated in Figure 2. Testing begins with the examinee choosing the difficulty level of the first item, at which point an item from the chosen stratum is drawn (without replacement) in a random fashion and administered. After this item is answered, the examinee is then asked to choose the difficulty level of the next item. This procedure continues until a predetermined number of items has been administered or a desired precision of proficiency estimation has been reached. After item administration is completed, the examinee's test performance is calculated using an IRT-based proficiency estimation
method. As with a CAT, because proficiency estimation is IRT based, the invariance property insures that the test performances of different examinees receiving a SAT can be directly compared even though they may have chosen to be administered tests that varied substantially in difficulty. Successful implementation of a SAT is largely dependent on the instructions presented at the beginning of the test. It must be explained to examinees that their test performance will be evaluated on the basis of the difficulty levels they choose as well as the number of items that they pass. Because most examinees are used to taking tests where performance is based solely on how many items are passed, examinees taking a SAT may tend to choose low difficulty levels unless adequate instructions are provided. Hence it is very important to provide examinees with clear instructions when administering a SAT. An example of instructions used with a SAT are found in Wise, Plake, Johnson, and Roos (1992).

The research on SAT conducted thus far has focused on its effects on test performance and its relationship to examinee affective variables. Several studies have compared SAT with CAT, finding that examinees receiving a SAT obtained significantly higher mean proficiency estimates (Roos, Plake, & Wise, 1992; Wise et al., 1992; Vispoel & Coffman, 1994). Moreover, the difference in mean estimated proficiency between SAT and CAT has been found to interact with other variables. Significant interactions have been found between test type and examinee scores on the Test Anxiety Inventory (Spielberger, 1980), with the difference in mean estimated proficiency between SAT and CAT increasing with examinee test anxiety (Rocklin & O'Donnell, 1991; Vispoel & Coffman, 1994; Vispoel, Rocklin, & Wang, 1994; Vispoel, Wang, de la Torre, Bleiler, & Dings, 1992). In addition,

Figure 2. Item Difficulty level strata in self-adapted testing.
Vispoel et al. (1994) found a significant interaction between examinee verbal self-concept and test type, with the largest difference in mean estimated proficiency between SAT and CAT being associated with low examinee verbal self-concept.

There also is evidence that the use of a SAT moderates the relationship between examinee anxiety and test performance. In two studies comparing SAT and CAT it was found that examinees administered a SAT reported significantly lower post-test state anxiety than examinees administered a CAT (Roos et al., 1992; Wise et al., 1992). It has also been found that a SAT yields proficiency estimates that are less related to test anxiety than those obtained when a CAT or a conventional test is used (Rocklin & O’Donnell, 1991; Vispoel & Coffman, 1994; Vispoel et al., 1994; Vispoel et al., 1992). The findings from these studies suggest that use of a SAT reduces the influence of anxiety on test performance.

CONTROL, MONITORING, AND SELF-ADAPTED TESTING

Although previous research on SAT has focused on its effects on anxiety and test performance, a SAT also affords an opportunity to measure elements of metacognition. To understand this, it is useful to consider the activities of the examinee during his/her test. A difficulty level is chosen by the examinee, an item is administered, the examinee answers the item, and the examinee is provided a choice for the difficulty level of the next item. This sequence is repeated until the test is completed.

We have observed that most examinees vary their difficulty level choices during the course of a SAT. Moreover, it has been found that many examinees tend to adjust their difficulty level choices to receive items that are well-matched to their proficiency levels (Wise et al., 1992). That is, many examinees taking a SAT appear to be motivated to attain the same difficulty-proficiency match that is explicitly sought by the computer algorithm in a CAT.

What psychological processes might be involved in attaining this match? We contend that two key processes are monitoring and control. Monitoring is required to assess the difficulty of the previous item, and to compare its difficulty to one’s perceived proficiency. Control is then required to make a strategic choice, regarding the next item’s difficulty, on the basis of the perceived degree of match between item difficulty and proficiency. If the match is sufficiently close, then the examinee will likely choose the same difficulty level as the previous item. If the match is not judged to be close then the examinee will change difficulty levels in order to attain a closer
difficulty-proficiency match. For example, if the examinee's monitoring process yields a judgement that the previous item was too easy, then the control process will choose a more difficult next item.

Thus, whereas most of the previous research on SAT has focused on the outcomes of taking a SAT, there is important information to be gained by studying the process of taking a SAT. Through an analysis of the SAT experience, we see that, although both monitoring and control play a major role in the examinee's strategic choices, the observable examinee behavior (difficulty level choice) most directly reflects the control process. Later in this chapter we outline several ways of using the data from a SAT to construct measures of the control process.

Some Methodological Advantages of SAT

Self-adapted testing provides a unique, unobtrusive method for gathering information about metacognitive processes, and especially of control. Indeed, examinees need not be given specific instructions about control or monitoring, or even know that their test behaviors provide relevant information about these processes. The fact that control processes are studied unobtrusively has two important advantages. One is that examinees are able to focus all of their resources on the test, rather than dividing their attention between performance and control-assessment activities. A second advantage is that direct measures of metacognitive control are available (i.e., item selection time and accuracy), rather than an indirect, subjective assessment of control (i.e., confidence or accuracy judgments).

SAT has a number of other advantages as well that pertain specifically to the task, test, and person constraints described earlier in this chapter. The most important of these is examinee control. Whereas all previous studies have asked examinees to complete a test designed by researchers, SAT enables an examinee to choose items that he or she feels are best suited to his or her proficiency without compromising comparability among examinees. With respect to the study of metacognition, individuals with a high degree of metacognitive control should be able to select difficult, yet answerable items. Those with less control may select test items that are less appropriate for them. Those with poor control may regularly select items that are too easy or too difficult. The self-controlled nature of SAT enables researchers to study the relationship among selection time, accuracy, and overall test proficiency, as well as a variety of self-report judgments made prior to, during, or subsequent to the test. Experiments could be expanded to examine motivational variables as well.
SAT also may increase the construct validity of proficiency estimates, and presumably measures of metacognitive control, by reducing confounds due to anxiety (Rocklin et al., 1995; Wise, 1994) and test difficulty. This helps to reduce or eliminate many of the test-based constraints typical of previous studies. For example, given that individuals select test items from a pool of calibrated items, the difficulty of these items should have little effect on the accuracy of metacognitive control. This is in stark contrast to traditional paper-and-pencil tests in which examinees monitor their performance with greater bias as test items increase in difficulty.

Another strength of SAT is the property of invariance, which enables each examinee to select items that are optimally suited to his or her proficiency. Differences in the absolute difficulty of items need not compromise estimates of metacognitive control. This means that measures of metacognitive control are comparable on the same scale even though individuals may be administered different test items and even though individuals differ with respect to underlying ability. The fact that SAT yields comparable estimates of proficiency and metacognitive control regardless of differences in ability eliminates a crucial person constraint in the study of metacognitive processes. It is likely that prior knowledge also has less impact on proficiency and control estimates than it would using paper-and-pencil tests. Although prior knowledge may greatly affect which items an examinee selects, item selection in itself does not affect estimates of proficiency. On the other hand, it is possible that individuals with no prior knowledge, or a great deal of it, may be poorly suited to the test if there are an insufficient number of test items near their true proficiency level.

In summary, we believe that self-adapted testing provides a unique opportunity to study on-line metacognitive control processes in an unobtrusive manner. The ability to do so permits researchers to explore a number of theoretical relationships among control, monitoring, and other cognitive skills (e.g., working memory span) that remain unanswered. We describe several intriguing questions in a subsequent section on future research. First we describe two methodological constraints on the use of self-adapted testing, then we describe a number of direct or derived measures of metacognitive control that are available from a typical SAT testing session.

Two Methodological Considerations

Two key issues must be addressed when using a SAT to measure metacognitive control strategies. First, the distribution of the item difficulties should span the range of the distribution of examinee
proficiencies, with enough items throughout the range that an examinee could take an entire SAT consisting of items with the same general level of difficulty. Having an item pool that is both "wide" and "deep" prevents examinees from being administered items that are not well matched to proficiency solely because well-matched items are unavailable.

A second consideration concerns the instructions given to the examinees. Without instructions for examinees to try to attain a close difficulty-proficiency match, it is unclear whether examinees who did not choose closely matched items did so because they were unable to match well or because they chose their items to attain another goal (e.g., reduction of test anxiety). Hence, examinees should be explicitly told to try to attain a close match. This, however, raises a troublesome new problem—how does one word such instructions such that examinees unequivocally understand their task?

The resources required to administer a SAT pose a third restriction on its use in metacognitive research. To administer a SAT, one must have (a) an item pool that is of sufficient size and has a broad range of item difficulty, (b) IRT parameter estimates for each item, and (c) computer software for administering computer-based tests. Regarding the item pool, it is important to have a distribution of item difficulties that spans the range of examinee proficiencies, and is "deep" enough that an examinee could choose a difficulty level that reflects a close difficulty-proficiency match many times without exhausting the difficulty level and being forced to receive items that are less well-matched. As an illustration, if a researcher plans to use eight difficulty levels in administering a 20-item SAT, the item pool should contain at least 160 items. Furthermore, IRT item parameter estimation requires a sizable calibration sample. Depending on the IRT model used, the typical recommendations for minimum calibration sample size range from 200 to over 1,000 examinees. Finally, special microcomputer software is needed to administer the SAT, such as the MicroCAT Testing System (Assessment Systems, 1994). Roos, Wise, Yoes, and Rocklin (in press) describe the program code needed to administer a SAT on the MicroCAT system.

QUANTIFYING METACOGNITIVE PROCESSES USING SELF-ADAPTED TESTING

Although both monitoring and control processes appear to be at work in a self-adapted test, the control process is more easily quantified using measures obtained during the testing process. A self-adapted test that is administered using computer-based testing
software such as MicroCAT (Assessment Systems, 1994) can provide a variety of information that is relevant to the measurement of metacognitive activities. When a MicroCAT test is administered, an output file for each examinee is created containing a detailed record of the examinee’s testing session. The file can contain an item-by-item record of the difficulty level chosen, whether the item was answered correctly or incorrectly, the examinee’s current proficiency estimate, and its standard error, as well as the time taken both to choose the item difficulty level and respond to the administered item. This information is readily obtainable from the MicroCAT testing system and does not require extensive programming skills on the part of the researcher. A guide to developing self-adapted tests on MicroCAT is provided by Roos, Wise, Yoes, and Rocklin (in press).

It is important to note that these measures are obtained in an unobtrusive manner. This mode of data collection allows examinees to focus their attention entirely on the test, alleviating concerns regarding the effects on test performance of requesting examinees to provide self-reports of metacognition.

There are several ways to quantify the relationship between the metacognitive control process and test performance (i.e., proficiency). To further illustrate these quantifications, we will refer to Tables 1 and 2. Table 1 is an example of a testing session for an examinee with a good match between proficiency and item difficulty, whereas Table 2 provides an example of an examinee with a poor proficiency-item difficulty match. Each examinee is administered 20 items from a pool of calibrated items that are partitioned into six mutually exclusive difficulty levels. For each item administered, the difficulty level chosen is displayed in the second column where Level 1 contains the easiest items and Level 6 contains the most difficult items. The difficulty parameter of the administered item is displayed in the third column. The difficulty parameters of the items are obtained using IRT estimation methods; these parameter values are matched to the scale of examinee proficiency, which typically has a mean of zero and a standard deviation of one. The higher the item difficulty parameter value, the more difficult the item. The fourth column indicates the difference between the examinee’s proficiency and the difficulty of the item. For example, the examinee in Table 1 had a final (i.e., end-of-test) proficiency estimate of -1.31, which is relatively low. The first item administered had a difficulty of -1.39, which indicates a close proficiency-difficulty match (.08). The fifth column lists the absolute value of the proficiency minus difficulty difference. The final column indicates the correctness of the examinee’s answer to the item.
Table 1. Testing Session for an Examinee With a Good Match Between Proficiency and Item Difficulty (Proficiency = -1.31, Standard Error = .319)

<table>
<thead>
<tr>
<th>Item</th>
<th>Difficulty Level Chosen</th>
<th>Item Difficulty Parameter</th>
<th>Proficiency - Difficulty Difference</th>
<th>Absolute Difference</th>
<th>Item Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>-1.39</td>
<td>0.08</td>
<td>0.08</td>
<td>Right</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>-1.13</td>
<td>-0.18</td>
<td>0.18</td>
<td>Wrong</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-1.73</td>
<td>0.42</td>
<td>0.42</td>
<td>Right</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>-1.42</td>
<td>0.11</td>
<td>0.11</td>
<td>Wrong</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>-1.28</td>
<td>-0.03</td>
<td>0.03</td>
<td>Right</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>-1.30</td>
<td>-0.01</td>
<td>0.01</td>
<td>Wrong</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>-1.32</td>
<td>0.01</td>
<td>0.01</td>
<td>Right</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>-1.50</td>
<td>0.19</td>
<td>0.19</td>
<td>Right</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>-1.25</td>
<td>-0.06</td>
<td>0.06</td>
<td>Right</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>-1.39</td>
<td>0.08</td>
<td>0.08</td>
<td>Right</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>-1.77</td>
<td>0.46</td>
<td>0.46</td>
<td>Right</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>-1.67</td>
<td>0.36</td>
<td>0.36</td>
<td>Right</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>-1.30</td>
<td>-0.01</td>
<td>0.01</td>
<td>Wrong</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>-1.57</td>
<td>0.26</td>
<td>0.26</td>
<td>Right</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>-1.27</td>
<td>-0.04</td>
<td>0.04</td>
<td>Right</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>-1.64</td>
<td>0.33</td>
<td>0.33</td>
<td>Right</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>-0.73</td>
<td>-0.58</td>
<td>0.58</td>
<td>Wrong</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>-1.24</td>
<td>-0.07</td>
<td>0.07</td>
<td>Wrong</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>-1.12</td>
<td>-0.19</td>
<td>0.19</td>
<td>Wrong</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>-1.14</td>
<td>-0.17</td>
<td>0.17</td>
<td>Wrong</td>
</tr>
</tbody>
</table>

Mean Over the Last 5 Items: -0.14 0.27
Table 2. Testing Session for an Examinee With a Good Match Between Proficiency and Item Difficulty (Proficiency = 0.32, Standard Error = .620)

<table>
<thead>
<tr>
<th>Item</th>
<th>Difficulty Level Chosen</th>
<th>Item Difficulty Parameter</th>
<th>Proficiency – Difficulty Difference</th>
<th>Absolute Difference</th>
<th>Item Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>-1.13</td>
<td>1.45</td>
<td>1.45</td>
<td>Right</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>-1.39</td>
<td>1.71</td>
<td>1.71</td>
<td>Right</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-1.73</td>
<td>2.05</td>
<td>2.05</td>
<td>Right</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>-0.73</td>
<td>1.05</td>
<td>1.05</td>
<td>Right</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>-1.42</td>
<td>1.74</td>
<td>1.74</td>
<td>Right</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>-1.24</td>
<td>1.56</td>
<td>1.56</td>
<td>Right</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>-1.12</td>
<td>1.44</td>
<td>1.44</td>
<td>Right</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>-1.14</td>
<td>1.46</td>
<td>1.46</td>
<td>Wrong</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>-1.28</td>
<td>1.60</td>
<td>1.60</td>
<td>Right</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>-1.30</td>
<td>1.62</td>
<td>1.62</td>
<td>Right</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>-1.32</td>
<td>1.64</td>
<td>1.64</td>
<td>Right</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>-1.50</td>
<td>1.82</td>
<td>1.82</td>
<td>Right</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>-1.25</td>
<td>1.57</td>
<td>1.57</td>
<td>Right</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>-1.39</td>
<td>1.71</td>
<td>1.71</td>
<td>Right</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>-0.54</td>
<td>0.86</td>
<td>0.86</td>
<td>Right</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>-1.77</td>
<td>2.09</td>
<td>2.09</td>
<td>Right</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>-1.67</td>
<td>1.99</td>
<td>1.99</td>
<td>Right</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>-1.30</td>
<td>1.62</td>
<td>1.62</td>
<td>Right</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>-1.57</td>
<td>1.89</td>
<td>1.89</td>
<td>Right</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>-1.27</td>
<td>1.59</td>
<td>1.59</td>
<td>Right</td>
</tr>
</tbody>
</table>

Mean Over the Last 5 Items: 1.84 1.84
The first measure of the relationship between metacognitive control and test performance is provided by the proficiency-difficulty differences. If instructed to attain a close proficiency-difficulty match, examinees should proceed through the test, monitoring the difficulty of the items administered and attempting to control subsequent difficulty level choices to attain a close proficiency-difficulty match. The degree to which an examinee is successful in controlling item difficulty will be reflected by the magnitude of his/her proficiency-difficulty difference at the end of the test, with smaller differences indicating greater control. Because the items are typically arranged randomly within each difficulty level, perhaps a more reliable index of the proficiency-difficulty match is provided by the average difference taken over the last five items. This is a measure of bias—the degree to which an examinee tends to select items that are too easy or too difficult. For example, the examinee in Table 1 showed a very good proficiency-difficulty match (-.14), whereas the examinee in Table 2 exhibited a poorer match (1.84) indicating a bias towards choosing item difficulties that were too low.

Another measure of control is provided by the absolute value of the proficiency-difficulty difference. This is an index of accuracy—the degree to which selected item difficulties are matched to an examinee’s proficiency. This index is also quite different for the examinees in Tables 1 and 2, with the examinee in Table 1 exhibiting a substantially more accurate match.

The standard error of the final proficiency estimate provides an alternative measure of accuracy. The more consistent the examinee is in choosing items well-matched to his/her proficiency level, the smaller the resultant standard error. Hence, the magnitude of the standard error indicates the accuracy of the examinee choices. The standard error for the examinee in Table 1 (.319) is substantially smaller than that for the examinee in Table 2 (.620).

Additional information is available from the testing session that may also prove useful in the study of control and (possibly) monitoring processes. One general type of information available is response latency; that is, the amount of time examinees take to (a) choose item difficulty levels and (b) answer items. Measures of this sort are very difficult to obtain in a traditional paper-and-pencil test but are easily and unobtrusively obtained when a test is administered via computer.

Researchers also may gain a better understanding of the control process through an investigation of the strategies used by examinees in selecting item difficulties. A computerized adaptive test provides an efficient model of control because the item selection algorithm
strives for a close proficiency-difficulty match. It would be particularly interesting, for example, to identify examinees who behave nearly as efficiently (or perhaps even more efficiently) as the computerized adaptive algorithm.

**DIRECTIONS FOR FUTURE RESEARCH**

Self-adapted testing allows researchers to investigate at least six questions pertaining to metacognitive control, and the relationship between control and monitoring, that have not been addressed adequately in previous research. We present these questions beginning with the most obvious and specific ones, gradually moving toward broader, more theoretical concerns.

Question one pertains to the relationship between metacognitive control and test performance. Researchers often assume that more accurate control leads to better test performance. SAT enables one to test this relationship directly while eliminating confounds due to item difficulty and presumed underlying ability. Existing theory also predicts a strong relationship between the accuracy of control judgments and performance (Nelson & Narens, 1994; Schraw, 1994). Researchers could study the impact of practice, domain familiarity, instructions, and other test-specific constraints via direct manipulation of these variables. Similarly, person-related variables such as prior knowledge and working memory span could be examined via blocking procedures, or treated as covariates.

Question two addresses the relationship between metacognitive control and response latency variables, including item selection and item response times. It is important to note that measures of response latency do not provide pure measures of a single cognitive activity per se. For example, item selection times, especially in the middle and later parts of a test, reflect some mix of control, monitoring, and performance processes. Nevertheless, SAT provides the best available methodology for assessing the relationship between control accuracy and response time. There is little theoretical precedent thus far regarding the relationship between control mechanisms and latencies. In general, response time and performance are related inversely, although the magnitude of the relationship, as well as its direction, depends on the type of variables being compared (Meyer, Irwin, Osman & Kounios, 1988). We expect a similar relationship between control accuracy and item response times. It is unclear, however, how control accuracy and item selection times are related. One plausible scenario is that individuals with a high degree of metacognitive control need little time to make strategic decisions, in part, because
many of these decisions are automated. This should lead to an inverse relationship between selection time and accuracy; that is, as control increases, selection times decrease. On the other hand, if item selection times include monitoring processes carried over from the previous item, we would expect a negative relationship between selection time and control accuracy. This assumes that monitoring is a relatively nonautomated, time-consuming process.

Data collected from SAT studies can be used to test competing hypotheses about the relationship between selection and response times, and control accuracy. One possibility is that this relationship changes systematically as a function of examinee knowledge, proficiency, practice, or test efficacy (Rocklin et al., 1995; Wise, 1994). These changes could be studied easily by blocking examinees on any of these variables or by manipulating controllable variables (e.g., instructions) directly.

Question three pertains to the specific relationship between expertise and control processes. Opinion appears to be split on this matter. Some researchers have suggested that monitoring accuracy is largely a by-product of domain-specific expertise (cf. Glaser & Chi, 1988). However, a number of recent studies (Glenberg & Epstein, 1987; Morris, 1990; Schraw & Roedel, 1994) failed to show a relationship between monitoring accuracy and domain expertise. It is important to note, however, that the relationship between monitoring and expertise may be quite different than the relationship between control processes and expertise. Currently, we know of no study that examines control accuracy across different levels of expertise.

SAT provides a format for investigating the relative impact of expertise on control processes, including performance accuracy, control accuracy, and item selection and response times. Although we would expect expertise to be positively related to test performance and estimated proficiency, we would not necessarily predict a corresponding increase in control accuracy. This reflects our view that control processes are, in part, domain-general phenomena (cf. Schraw, Dunkle, Bendixen, & Roedel, 1995). Although expertise should enable examinees to perform better on a test, their expertise need not improve their ability to control or monitor with a high degree of accuracy.

A fourth question is the degree to which control accuracy is related to other cognitive variables such as general aptitude and working memory span. Very little research has been done in this area in general. Of studies that have investigated these relationships directly or indirectly, there is little evidence that aptitude is related
strongly to metacognitive processes in children (Alexander, Carr, & Schwanenflugel, 1995; Swanson, 1990) or adults (Pressley & Ghatala, 1990; Yan, 1994). We know of no study investigating the relationship among control and monitoring accuracy and traditional indices of the speed and accuracy of working memory.

Research in this area is important for two reasons. One is to establish the degree to which metacognitive processes such as control and monitoring are related to "hard-wired" cognitive differences such as general intelligence and working memory capacity (cf. Jensen, 1992). Evidence that metacognition is not related strongly to these variables would highlight the flexible, developmental nature of metacognitive knowledge. A second reason is to examine the compensatory relationship between measures of cognitive ability and metacognitive knowledge. In a ground-breaking study by Swanson (1990), for example, metacognitive knowledge contributed to complex problem solving among young adolescents over and above the effect of ability. This finding suggests that metacognition may follow a separate developmental path, and may act independent of other cognitive mechanisms (see Alexander et al., 1995, for a further discussion).

Question five pertains to the still elusive relationship between control and monitoring processes. Several theorists have distinguished clearly between control and monitoring processes (Koriat, 1994; Nelson & Narens, 1990; Pressley & Ghatala, 1990).

Nevertheless, much of the empirical literature in the field has focused on monitoring rather than control processes, due in large part to the difficulty researchers face when measuring control.

Some believe that control and monitoring are practically, if not statistically, linked (Nelson & Narens, 1990, 1994). Others believe that monitoring is both functionally and statistically independent of control, and in fact, represents a fundamentally different type of cognitive activity (Koriat, 1993, 1994).

The literature is in need of further contributions on this point. We believe self-adapted testing methods can be used with tremendous advantage to address this question. Previously, we described how control processes may be quantified in a SAT via direct and indirect measures obtained unobtrusively. It also is possible to obtain measures of monitoring via subjective judgments made after answering a test question within the otherwise computer-based format of SAT. Control and monitoring indices could be compared over the course of a test to determine their relationship. If the two are linked, one would expect monitoring judgments made at item selection to be linked to
item selection at item $i + 1$. Data of this type, as well as a variety of derived indices of control and monitoring, could be used to test the efficacy of a regulatory loop that connects monitoring and control functions. In this view, monitoring processes provide data-driven feedback to control processes that use this feedback to iteratively guide future performance. This presumes that monitoring and control processes are flexible, reciprocal processes that communicate with each other, even if they do not share a common set of cognitive resources.

It is possible that control and monitoring processes are related in different ways as a function of expertise. For example, control and monitoring may be related more strongly as expertise increases, provided these processes become mutually encapsulated within the expert domain (Glaser & Chi, 1988). If control and monitoring skills remain domain-general in nature, then expertise within a domain should not matter. Another possibility is that control and monitoring are unrelated (Koriat, 1993, 1994). In this view, monitoring processes are “parasitic” in that they are based on domain knowledge and efficacy beliefs within the domain, rather than a metacognitive mechanism that actually monitors the accuracy of performance independent of domain knowledge.

A final question addresses the degree to which individuals are better able to control their subsequent performance than, for instance, a minimum-error computer algorithm. Part of our interest in this question stems from the finding that some individuals perform better on a SAT than on a comparable CAT (Rocklin, 1994; Wise et al. 1992; Wise, Roos, Plake, & Nebelsick-Gullett, 1994). Wise (1994, p. 18), for example, stated “when examinees are allowed to choose their test item difficulty levels, they perceive a sense of control over the test, which serves to reduce anxiety” and which presumably improves performance. Echoing Wise’s (1994) thoughts on perceived control, Rocklin et al. (1995, p. 114) stated that “the effects of self-adapted testing can be attributed specifically to the control that examinees exert over the difficulty levels of items they attempt.” One explanation of the difference between SAT and CAT versions of the same test is that many examinees experience less anxiety when taking a SAT (Wise et al., 1994). Another explanation, although not mutually exclusive from the reduced anxiety hypothesis, is that some individuals are better able to control their performance than even the most accurate computer-driven selection algorithms. One way to test this difference is to offer good and poor controllers the opportunity to take similar exams using both SAT and CAT formats. Coupled with on-
line or retrospective verbal reports, a comparison between the two methods may illuminate some of the subtle control processes used during testing.

These six questions present an impressive array of topics that warrant further research. Understanding control processes with more precision is important in and of itself. However, understanding the crucial relationship between control and monitoring is even more important, because it is inconceivable that researchers could claim to understand metacognition without understanding the locus and functions of control and monitoring under a wide variety of circumstances, as well as the relationship between them. Similarly, it is essential to understand what makes a highly metacognitive person so able to self-regulate his or her behavior. Comparing good and poor controllers (and monitors) to existing computer software may provide some illustrative insights that increase our understanding, while posing new research questions.

SUMMARY AND CONCLUSIONS

This chapter explored some of the possibilities of using a computer-based testing format to investigate metacognitive processes. We reviewed recent research on control (i.e., regulatory processes used to guide cognitive activities) and monitoring (i.e., regulatory processes used to evaluate the present success of one's performance) functions of metacognition.

After highlighting some of the basic assumptions of computer-based testing, we described several specific strengths of self-adapted testing (SAT). We argued that SAT alleviates a number of serious methodological problems endemic to traditional tests. These included confounds due to differences in ability, prior knowledge, and item difficulty. A more salient problem was that traditional tests do not allow examinees to exert full control over their test-taking behavior. SAT eliminates this problem, and simultaneously offers researchers the opportunity to gather valuable information unobtrusively.

We next considered some of the direct (e.g., item selection time) and indirect (e.g., control accuracy) measures available when using SAT. These measures can be used to answer a host of questions about metacognitive control, as well as the relationship between control and monitoring processes. In addition, it is possible to compare good and poor monitors, as well as to compare the same examinee under CAT and SAT testing conditions. These comparisons offer a unique opportunity to study many aspects of metacognition in a much more direct, yet unobtrusive manner.
Our main conclusion was that computer-based testing formats offer a number of new methodological avenues for the study of metacognition. We proposed six questions that warrant considerable research over the next decade. Chief among these is the relationship between control and monitoring processes, whether these processes share a common pool of resources, and whether they enjoy a reciprocal exchange of information indicative of a regulatory loop. Although little was said concerning developmental issues, we see little difficulty applying these procedures to younger examinees, provided individuals have some knowledge of the test domain, and researchers have access to a calibrated pool of test items.

Finally, despite the tremendous potential of self-adapted testing as a tool for measuring metacognition, we wish to emphasize its essential compatibility with other measurement techniques. SAT seems amenable to on-line and retrospective verbal reports, as well as to on-line subjective performance judgments similar to those used in most monitoring studies. SAT also provides an opportunity to investigate the criteria examinees use to select test items. Concurrent verbal reports may be highly valuable in this regard.

REFERENCES


Schunk (Eds.), Self-regulated learning and academic achievement: Theory, research, and practice (pp. 169-200). New York: Springer-Verlag.


I am going to begin with claims that may seem heretical at the Buros Institute, the host for this symposium: Much can be understood about cognition and its metacognitive regulation through qualitative analysis. Qualitative analyses of complex cognitive and metacognitive processes makes a great deal of sense before even attempting quantitative analyses of those processes. In particular, I am going to explain here the advances made by my associates and me in understanding skilled reading using the method of constant comparison, a qualitative approach for developing what Strauss and Corbin (1990) refer to as grounded theories. If that does not offend Buros regulars, perhaps the types of data used as input to the theory construction process will. I believe, as do others (see Ericsson & Simon, 1993), that the most telling analyses of complex, conscious, self-regulated cognitive processes have been produced using verbal protocol procedures—that is, when people have thought aloud as they performed complex tasks. My associates and I have been using verbal protocols of reading to develop grounded theories of consciously regulated reading.

Given that preconceptions do influence research, it is important to lay out one’s assumptions and understandings about a to-be-researched problem at the outset of the study and to audiences who
must evaluate the work. Thus, I begin this chapter by laying out briefly my theoretical sensitivities before I conducted the research reported here, an essential step in qualitative analyses. They included a history of success with both verbal protocol analyses and grounded theory approaches, a long-term interest in reading comprehension, some successes in studying it using traditional quantitative, experimental approaches, but also some important frustrations doing so.

After laying out my preconceptions, I will cover a specific verbal protocol study conducted by my colleagues and me in which constant comparison was used to develop a grounded theory of how social sciences professors read research articles in their areas of expertise. This will be followed by a discussion of how Pressley and Afflerbach (1995) used constant comparison to generate a more general grounded theory of the conscious processes in reading. They used all of the data generated in verbal protocol studies published to date, with the result a grounded theory that is a qualitative meta-analysis of the verbal protocols of reading reported to date. I will conclude the chapter with a brief discussion of the implications of the work reviewed here for future individual research projects on conscious processing during reading, the development of standardized measures of reading comprehension skill, and the construction of more complete grounded theories of complex cognition and behavior.

BACKGROUND: MY THEORETICAL SENSITIVITIES AS I EMBARKED ON THIS RESEARCH

Beginning with my earliest research, I have been interested in the strategies people use to accomplish academic tasks, and in my second year of graduate school I discovered the power of verbal protocols to reveal cognitive processes, many years before Ericsson and Simon’s (1984) book made the approach more respectable in the eyes of many research psychologists. In particular, Pressley and Levin (1977) asked students in grades 5 through 9 to talk aloud as they studied paired associates in anticipation of a memory test. We demonstrated in that work a clear developmental shift in the likelihood that students would verbally elaborate paired associates as they studied—that is, embed word pairs in a meaningful sentence. Especially striking was that whether and how much students reported elaborating was a much better predictor of objective memory performance than age, a satisfying outcome for us at the time given our theoretical conviction that Jenkins (1974) was correct, memory depends much more on what one does to remember than on other factors. Pressley and Levin
were widely cited and this was considered one of the best contributions I made in the area of memory development. The acceptance of that work did much over the years to convince me personally of the value of the think-aloud approach.

Since my graduate school days, I also have been interested in how cognitive strategies can and do mediate learning of text (Pressley, 1976, 1977). Although most of my research in the decade following graduate school was not concerned with text processing, I returned to the study of comprehension processing in the late 1980s. Using a variety of qualitative methodologies, including ethnographies, interview studies, and case studies (e.g., Pressley et al., 1992) my group made fairly rapid progress in our efforts to understand how elementary students can be taught comprehension strategies. That qualitative methods produced rapid understanding of high quality comprehension instruction, work that has been applauded in many ways, fueled my enthusiasm for qualitative research.

As my students and I tackled the problem of comprehension strategies instruction in the elementary grades, we also began research on adults' naturalistic processing of text. On the positive side, we had some success using quantitative methods to document that college students often are not aware whether they have learned text content they have studied (Pressley, Snyder, Levin, Murray, & Ghatala, 1987) or even when they have completely missed the point of something they have read (Pressley, Ghatala, Woloshyn, & Pirie, 1990a, 1990b). That is, we were able to document somewhat surprising monitoring failures in adult readers—surprising, at least, from the perspective of those who believed that comprehension monitoring failures were more a problem of childhood than adulthood (Markman, 1981).

With the successes came some frustrations, however. For example, Barbara Snyder's master's thesis with me at Western Ontario involved a number of well-controlled, quantitative comparisons of students' overt behaviors as they read from a textbook in use in one of their courses. We presumed that this would be a window on the nature of skilled reading. That proved not to be true. The only conclusion that we could draw from the study was that college students do a great deal of beginning-to-end rereading as they study textbook chapters. Although this finding pointed out a disturbing quality of college student reading, and it is a result that has proven replicable and continues to disturb those who worry about the efficiency of college student study (Cordon & Day, 1995), it seemed to Barbara Snyder, me, and reviewers of the paper resulting from the study, that much more must be going on. I suspected that the
impressive text representational abilities of college students that had been documented by Kintsch (e.g., 1982, 1983, 1988, 1989), Graesser (e.g., 1981; Graesser & Bower, 1990), van den Broek (1990a, 1990b), and others were in part the products of conscious processes, processes that were essential for me to understand, given my career-long interest in how purposeful processing affects learning, memory, and comprehension.

I went with the instincts that had served me well in understanding paired-associate learning. I started inviting people to my office, asking them to think-aloud for me as they read. What I heard was a bit overwhelming. Reader after reader provided extremely rich think-alouds, ones filled with strategies, attempts to make inferences, and great intellectual activity in general, including reflection on and evaluation of what was read. As I reviewed the quantitative, experimental studies of text processing conducted in the 1980s, what was surprising to me was that none of these studies seemed to be capturing the richness of the processing that I heard readers describing. Moreover, as I read the think-aloud studies of reading conducted in the 1970s and 1980s, I had the same feeling, only more intense, for I knew the think-alouds I had witnessed in my office were filled with information about reader strategies, attempts to construct inferences, and reflections on text.

What was going on in the verbal protocol studies of reading? Most of the think-aloud studies were designed to test particular hypotheses—to determine if particular types of processing were occurring. That is, many of the investigators believed some particular type of processing was occurring in reading and conducted their think-aloud analysis to confirm such a possibility or elucidate the processing further. I realized what was needed was think-aloud studies in which the researchers were as open-minded as possible about the processes that might be reported. I was optimistic that such an approach might work in light of my recent success in studying elementary-level comprehension strategies instruction. By approaching that work with the goal of constructing as complete a grounded theory of teaching as possible, my colleagues and I had constructed a theory that included much that others had missed when they had studied elementary-level comprehension strategies instruction. I knew what I had to do: It was to apply such an open-minded, grounded theoretical approach to the analysis of verbal protocols of reading. Before describing the efforts of my colleagues and me to do this, I review briefly some essential prerequisite materials: how grounded theories are constructed and why there is reason to believe that verbal protocols of reading are valid indicators of reading processes.
Development of Grounded Theories Using the Method of Constant Comparison

Before a scientist will use a methodology, she or he must be convinced that it is rigorous and effective in doing what is claimed it does. Strauss and Corbin's (1990), Basics of Qualitative Research: Grounded Theory Procedures and Techniques particularly convinced me that construction of grounded theories made sense. I describe here some highlights of their perspectives on data and its analysis.

Construction of a grounded theory begins with collection of data. That is, this is an approach for developing a theory that is grounded in data. As a dyed-in-the-wool empiricist, this aspect of Strauss and Corbin's thinking was exceptionally appealing to me. The researcher who is attempting to develop a grounded theory may spend a great deal of time observing behaviors in a setting of interest, interviewing informants, or, particularly relevant here, have people think aloud as they do a particular task. Many types of data can inform the development of a grounded theory. In some cases the researcher will rely on only one type of data, in other cases on variations of one type of data (e.g., different types of verbal protocols of reading), and in still other cases, several types of data.

The task then is to induce regularities from the data collected, through a method known as constant comparison. Thus, the researcher goes through the data systematically looking for meaningful clusters and patterns—behaviors that seem to go together logically. It is then necessary to name the clusters, to come up with category names for the behaviors included in the clusters. Such an analysis often results in a number of categories.

The next objective is to attempt to identify evidentiary support for the categories. The investigator, however, is always open to—and actually looking for—data inconsistent with the emerging categories. This can be done by reviewing previous data, but typically also includes the collection of new data.

We note that qualitative researchers typically begin their analyses early in the data collection. As tentative categories emerge, there is opportunity with every new data collection to look for support or nonsupport of categories—to compare tentative conclusions with conclusions suggested by new data. The researcher may change categories or their names, delete categories, or add them in light of new data. In short, there is fluid interaction between data collection, data analyses, and construction of conclusions. Analyses and data collection are interwoven enterprises.
Eventually, there is a stable set of categories. The task then is to define the categories precisely, in terms of defining properties, and to begin to organize these categories in relation to one another. For example, categories can be placed in hierarchical arrangements with each category defined in terms of its defining properties. Once the categories have been identified, fully defined, and placed in hierarchical arrangement (with these categorizations, definitions, and arrangements challenged by checking against the data a number of times), the researcher can begin to feel that the data on hand are understood about as well as they are going to be understood. Data collection continues until no new categories, defining features of categories, or relationships between categories are being identified. This may take a while or it may happen fairly quickly. Strauss and Corbin (1990) are emphatic that researchers must continue to analyze the data—must continue to compare emerging conclusions against new data—until the point is reached when no new information is being generated, for to do otherwise results in an incomplete grounded theory.

Just as it is possible to evaluate the quality of quantitative studies, it is also possible to evaluate qualitative studies—on about the same dimensions. The language is different, however (Guba & Lincoln, 1982; Lincoln & Guba, 1985). Dependability is the qualitative analysts’ term for reliability. That is, the qualitative analyst must convince that most people would come to the conclusions that are drawn based on the sample of data analyzed. Rather than worrying about internal validity, qualitative researchers are concerned with credibility. To the extent that the case is strong that the grounded theory captures the reality of the situation studied, the greater the credibility of the study. Rather than external validity, the qualitative analyst values transferability—that the analysis was conducted in a setting representative of the universe to which the researcher wants to generalize. Confirmability is the term used instead of objectivity, with confirmability generally high when something like triangulation occurs in the study—when multiple indicators are used to buttress conclusions. The best qualitative studies are high on all of these characteristics.

Validity of Verbal Protocols of Reading

If verbal protocols of reading are to be used as indicators of skilled reading, it is essential that there be clear relationships between verbal reports of cognitive processes during reading and actual reading. As it turns out, the track record on this count is strong for adults readers, including the following outcomes:
Hare (1981) reported that good compared to weaker college student readers were more likely to monitor their comprehension as they read and set into motion fix-up strategies when comprehension was less than complete.

Olson, Mack, and Duffy (1981) observed correlations between self-reported strategies at particular points in text and the speed of processing at those points. That is, reading was slower early in a story when readers reported storing background information presented in the story and formulating hypotheses about the stories read. At points where substantial inferential activities were reported, processing was slower. Subjects reported simply confirming their suspicions as they finished text with relatively rapid reading times near the end of text.

In Trabasso and Suh (1993), self-reported inferential activities predicted a variety of performance measures related to the inferences, including reading times and long-term retention of stories.

Wade, Trathen, and Schraw (1990) examined the overall patterns of strategy use reported by college students as they read The Sea Around Us. They detected six types of profiles of text processing, varying from ones that reflected extensive responding to text to minimal responding. By far, one of Wade et al.'s groups was more sophisticated in their strategies use than any of the other five. This group, which Wade et al. (1990) referred to as "good strategy users," following a categorization suggested by Pressley, Borkowski, and Schneider (1987), was more diverse in their strategic responses to text than were other participants in the study. They made notes, paraphrased, outlined, and/or constructed diagrams as they read. They varied their reading speed from skimming to slowing, and they reread when it was necessary. They made use of their notes and mental notings to review the text read after reading. With respect to recall of important information in the text, there was more than a half standard deviation recall difference favoring the good strategy users relative to the next best group.

Guthrie, Britten, and Barker (1991) reported that the strategies self-reported by college students as they searched documents for information correlated with how efficiently they searched text.

Haas and Flower (1988) observed that graduate students were more likely than undergraduate students to do "rhetori-
cal reading” of a section of an undergraduate textbook. That is, they reported attempting to understand the author’s intentions in writing the text as it was written.

- Deegan (1993) observed that first-year law students who were doing well in law school read differently than first-year law students experiencing difficulties in school. Specifically, the better students were more likely to respond to text with questions about the meaning and structure of a law-related text they read.

- Lundeberg (1987) reported that legal experts were more likely than legal novices to attend to important information in a legal case they read, and to overview the case, attempt to summarize it, evaluate it, and reread the case analytically.

- When Earthman (1989) had graduate students in English and freshmen read short stories and poems, she found that the graduate students were more likely than freshmen to work at filling in gaps in meaning in the texts and were more likely to relate texts to knowledge of the world. The graduate students were also more likely to take alternative perspectives while reading the literary works.

- Graves and Frederiksen (1991) observed considerable differences between the think-alouds of English professors reading an excerpt from *The Color Purple* and college sophomores doing so: The professors were more aware of the functions of the narrative in the text as well as the relationship of the author to the reader of the text. The experts viewed the text as the result of deliberate choices made by the author, with their perceptions of these choices affecting their understanding of the text.

- In Wineberg (1991), when historians read American history textbook material, they were much more likely than high school students to search for the authorial intentions and hidden meanings. High schools students treated the texts more as factual documents containing information that was not open to question. The historians questioned.

In short, a variety of investigators, collecting diverse think-aloud data as people read, have observed correlations between reported reading behaviors and reading performance or between reported reading behaviors and level of reading ability. Although my view at the outset of this study was that much more validating data would be desirable, I was struck that the studies validating verbal protocols were not countered in the literature by failures to obtain verbal
report-reading relationships. There is good reason to have confidence in the validity of verbal protocols of reading.

Summary

Since the earliest days of my career I have been aware of the verbal protocol approach. It worked well for me in the past, as it has for other investigators interested in complex cognition. I have also been interested in comprehension processes throughout my career. Although I enjoyed some research success in analyzing comprehension, I felt that much important about comprehension was not coming through either in existing experimental analyses or verbal protocol studies, an awareness developed as I listened to verbal protocols of text processing much richer than the descriptions of processing in the existing literature. My success with the grounded theory approach as a tool for the analyses of comprehension instruction impressed upon me the power of this approach to elucidate complex phenomena. Thus, it made sense to me to apply that approach to analyses of verbal protocols of reading, believing that if I did so, I would produce much richer descriptions of reading than had been generated in the past. What follows is a description of how my colleagues and I did so in a single study, followed by a brief review of how Afflerbach and I applied the method of constant comparison to 40 verbal protocols of reading to produce a general grounded theory of conscious processes during reading.

HOW SOCIAL SCIENCES PROFESSORS READ JOURNAL ARTICLES

What Wyatt, Pressley, El-Dinary, Stein, Evans, and Brown (1993) wanted to do was document reading at its best—what it might look like when exceptionally skilled readers are reading content that is interesting to them for a purpose that is important to them. As members of a Washington DC university community, a convenience sample of skilled readers was faculty members. Given what we knew about expertise (Chi, Glaser, & Farr, 1988), it seemed likely that the most sophisticated reading might be observed when readers read in domains in which they had high prior knowledge. Thus, we decided that we would ask professors to read in their areas of expertise. So that their interest would be high, and the material was being read for an authentic, meaningful purpose for the professor, we felt that it would make most sense to allow the readers themselves to select the articles they would read.
In identifying a pool of potential "expert" participants, the following criteria were applied: (a) The participant possessed a doctorate in a social or behavioral science. (As social scientists ourselves, my students and I felt that we would be in a better position to understand the think-aloud comments of social scientists compared to natural scientists, humanities scholars, or professors in some other field foreign to us.) (b) The participant had published at least five articles in selective outlets over the last 5 years. We felt that this criterion would assure a fairly select sampling within the select category of university professors. None of the participants, however, had written on reading strategies, nor was there any reason to believe that any had particular scholarly expertise about the nature of skilled reading.

Procedures

At an initial meeting, the participating reader was told that the purpose of the study was "to investigate how experts stay current in their fields of expertise." Participants were asked to select three research articles that they had not yet read but would be interested in reading as part of "staying current in their field." The researcher requested that the participant not begin reading the articles—that they make their selections on the basis of author and title only.

The entire second session was recorded on audio tape. At the start, the investigator explained that the session would be devoted to working with one of the articles. The subject then chose an article from the three she or he had identified. In all cases but one, the article was the report of original empirical research; in the outlying case, the article was a position piece on a particular research direction.

Participants were directed to "read the article as they normally would." They were encouraged to think aloud as they went through the article, offering any comments or explanations they wished. Using a duplicate copy of the article, the investigator noted the participant's actions, attentive to any aspect of the participant's behavior that pertained to processing of the article. For example, the researcher noted indications of the reading path taken through the article—when different sections were begun, pages turned, text underlined, verbatim statements made, and so on. Observations of participant's nonverbal behaviors also were noted on the researcher's copy. If more than 2 minutes passed without any verbal comment from the reader, the investigator prompted the reader with the question "What are you doing now?" At the end of the session, the researcher collected the participant's copy of the article so that any markings the participant made could be analyzed further.
Records of this second meeting were expanded into a comprehensive chronological description of the participant’s activities while reading the article. In this process, the researcher’s notes, audio recording, and any annotations on the participant’s copy of the article were combined to generate a thick description of the participant’s reading behavior.

In the third and final participant-investigator meeting, the investigator gave the participant a copy of a process description of their reading. This provided an opportunity for participants to identify problems in the description and analysis of their reading strategies. When the participant disagreed with the description or analysis (which was extremely rare and never with respect to a major conclusion in the protocol), the disagreement was noted and an adjustment in the protocol considered later by the researcher, following additional review of the raw data.

Analyses

For the first five readers in the study, five members of the research team each worked with a participant’s protocol and began an analysis of the observed reading behaviors, following a variation of the method of constant comparison (Strauss & Corbin, 1990) described earlier. In particular, they examined and reexamined a protocol, attempting to identify categories that exhaustively accounted for the behaviors in it. Then, these five researchers and a sixth member of the research team met and compared categories they had observed. Each researcher then reexamined the protocol they had analyzed in light of categories identified by the other researchers. Over the course of several meetings, analysis and discussion of strategies used by these first five participants resulted in a long list of individual strategic behaviors. Additional meetings then occurred, each one followed by reanalysis of the reading protocols of the first five participants and reflection on the categories of reading behaviors that typified what was observed in the reading of the first five participants. After about 8 weeks of reanalyses and reflection, the six co-investigators were satisfied that the most critical reading behaviors were captured adequately by the categories summarized in Table 1. The scoring categories were grouped into theory-based sets and subsets as reflected in the organization of Table 1.

The 10 protocols of reading subsequently collected also were scored in terms of the Table 1 criteria. Two researchers scored each protocol: the researcher who had had face-to-face contact with the participant and one other member of the research team. The team
Table 1

Linearity and Nonlinearity of Reading
- Reader either surveys text before reading it or does not.
- Reader either generally reads article from front to back or does not.
- Reader either reads large section of article in a linear fashion or does not.
- Frequency of jumping forward (jumps ahead to another section, staying at least 30 seconds) or looking forward in text for particular pieces of information (e.g., footnotes, results, references) and returns.
- Frequency of jumping back (Jumps back to another section, staying at least 30 seconds) or looks back in text for particular pieces of information and returns.
- Frequency of reading selectively in linear fashion (skips some information, then reads closely) during reading of the abstract ... introduction ... methods ... results ... discussion/conclusion ... references.

Goal Awareness
- Whether highly aware (before reading) of specific information being sought from the article and looking for it.
- Whether looking for information relevant to personal and/or professional goals (own research, writing, teaching, bibliography).

Awareness
- Frequency of reading aloud (and self-reports that he or she would read aloud if reading alone).
- Frequency of exploiting personal strengths (e.g., says can understand tables better than text, so more attention to tables, or vice versa).
- Frequency of closely attending to tables/figures.
- Frequency of talking about things, “I typically do when I read.”
- Frequency of varying reading style according to relevance of text to reading goals. (Style includes slowing for careful reading, skimming, and very fast skimming.)
- Frequency of expressing own biases/expectations toward text.
Planful

- Frequency of reported watching for particular information throughout reading.
- Frequency of reported decisions to continue reading (based on the abstract or something other than abstract).
- Frequency of claiming intent to read section in specific order.
- Frequency of adjusting attention to material depending on relevance to reading goals.
- Frequency of noting parts of text (e.g., references) to read later or to remember for future reference.

Monitoring

- Frequency of backtracking. (Rereads a sentence for clarification or backtracks for stated purpose of clarification.)
- Frequency of noting explicitly how difficult the text is to read (reading is easy, difficult, she/he does not understand the text, something in text is puzzling).
- Frequency of noting explicitly when something in text is worth or not worth noting.
- Frequency of noting explicitly when something in text is already known or not known to him/her.
- Frequency of noting explicitly when something is taken from another source (e.g., from a named researcher’s work).

Relating Information to Prior Knowledge Base

- Frequency of reading reference list to activate prior knowledge.
- Frequency of anticipating/predicting information that will be presented; testing predictions.
- Frequency of reacting to information based on own knowledge (including reactions to the author being read, others authors cited in the text, methods, analyses, content, discussion, or text structure of the paper).
- Frequency of reacting to text based on very personal prior knowledge (e.g., own theories, own writing, knows author personally).
- Frequency of noting that text contradicts a belief held by the reader.

Continued.....
Table 1 (continued)

Evaluative Reactions
- Frequency of evaluating relevance to goals.
- Frequency of evaluating whether what is being read is the specific information being sought from the article.
- Frequency of evaluating whether information is relevant to personal and/or professional goals (own research, writing, teaching, bibliography).
- Frequency of evaluating the text (including reactions to literature review, particular citations, theoretical perspectives, methods, analyses, and results—including the novelty of findings, conclusions, discussions, implications, writing/editing style, and biases of the author).

Going Beyond the Information Given (Elaborations)
- Frequency of constructing conclusions or summary interpretations beyond information provided in article. (Comes up with summary interpretation of results, tables, or discussion/conclusion.)
- Frequency of constructing paraphrases/explanations of what is in the text and/or gives examples.

Integration
- Frequency of going back and forth in text (to go to table or figures or to guide further reading in this article). Goes back and forth between figures/tables and text or compares figures/tables with one another to integrate.
- Frequency of getting information explicitly from text on figure or information from figure on side of text or side of figure.
- Frequency of verbally relating material from different parts of text.
- Frequency of summarizing the whole paper after reading it.
- Frequency of indicating she or he will be looking at other materials later with eye to relating to what is in this text.

Elucidation of Discourse Structure
- Frequency of mentioning division or relations among different parts of a section or marks major divisions of an argument (e.g., by writing brief title for division, numbering steps).

Continued.....
Table 1 (continued)

Written Responses

- Frequency of highlighting (frequent marking of text to highlight, including underlinings, check marks, arrows, brackets, boxes) and marking references/terms to find later.
- Frequency of elaborating (making brief summaries of text, including marginal notes); sketching the design of the study in writing; relabeling figures/tables; adding more information to figures/tables; rewriting some information in clearer, more memorable form.
- Frequency of writing notes on separate piece of paper or computer.

Affective Reactions

- Frequency of expressing positive affective reactions.
- Frequency of expressing negative affective reactions (including anger, tiredness, or boredom).
- Frequency of expressing interest.
- Frequency of expressing lack of interest.
- Frequency of expressing surprise.
- Frequency of using expletives or slang.

Nonverbal Responses

- Frequency of laughing, looking puzzled, gesturing, giving raspberry, scratching chin, putting hands on forehead.

continued to meet frequently to assess whether the existing set of reading strategies needed to be modified. There were very few behaviors (and no potentially important ones in our view) produced by the 15 participants that were not consistent with the Table 1 categories. When participants’ behaviors were categorized as never occurring, occurring once, occurring 2 to 4 times during the session, or occurring 5 or more times, there was little disagreement at all between raters (i.e., although the two raters might disagree whether 7 or 8 instances of a behavior occurred, this made no difference when the response classification was that the behavior occurred “5 or more times”).

Results

All 15 readers in the study were very active, using well-regarded comprehension strategies, such as predicting and verifying predic-
tions, summarizing, elaborating on text, seeking clarification, and reading selectively. The readers in this study monitored comprehension and important characteristics of the text, such as its difficulty level and relevance to reading goals. They evaluated the adequacy of text form and content. Indeed, each of the behaviors in Table 1 was evidenced by the majority of readers in this study. Moreover, there was nothing rigid about their articulation of strategies and prior knowledge, but rather, the readers were highly responsive to text, shaping their reactions on the basis of text content, its relationship to their prior knowledge, and their purposes in reading the text.

Summary

What Wyatt et al. (1993) reported was the most elaborate set of comprehension processes ever identified in a verbal protocol study. This was undoubtedly due partly to the sophistication of their readers and to the match between readers’ prior knowledge, interests, and what they were reading. I believe, however, that the detailed model of text processing emerging from Wyatt et al. (1993) also was due to the analytical approach taken. We reflected long and hard on what the readers were doing, reflecting and rereflecting on the possibilities of missed categories of responses. In the end, what we had was a high quality qualitative study. The results were dependable, in the sense that we believe most observers armed with the Table 1 criteria would have scored the individual readers as we did, given that the members of the research team were able to score the protocols with great reliability. With respect to credibility, there was simply nothing left over to score after the Table 1 criteria were applied: An internally valid model accounts for everything in the data, and this analysis did so. There was transferability. The scoring scheme developed with the first 5 readers generalized to the next 10 readers.

Even so, this study did not capture all the conscious processes that are skilled reading. No individual verbal protocol study could do so, for each is limited with respect to type of reader and the type of material being read. To capture all of the conscious processes that are skilled reading, what is required are diverse readers reading a variety of materials. More positively, when all of the verbal protocol studies of reading are considered, there is an enormous range of readers reading a great variety of materials. I realized that it might be possible to construct a general model of conscious processing by collapsing the outcomes obtained across these studies. I also recognized that the method of constant comparison could be adapted to do that, with the promise of a general grounded theory of conscious reading.
Wyatt et al. (1993) was simply one in a long line of verbal protocol studies of reading, beginning with Squire’s (1964) analysis of how teenagers process short stories. Across these studies, many different readers have been studied from grade-4 students to middle-school and high-school students to college students and their professors as well as other highly skilled professionals. Many different types of materials have been read from poems and paragraphs to short stories and expository pieces. Readers’ goals have varied from study to study. The specific instructions provided to readers varied as well. Thinking aloud was operationalized in a number of ways in these studies, from completely self-regulated thinking aloud during reading to reporting thoughts at designated points in text to reporting thoughts shortly after reading is completed. Sometimes the scoring of data was grounded completely in the data, as in Wyatt et al. (1993). More often, it was not, with investigators interested in particular processes and hence, scoring the protocols selectively. One reaction to all of this variability might be to throw up one’s hands and exclaim that general conclusions about reading could not possibly emerge from it. If the goal is to understand every conscious process that might occur during reading, however, the variability across studies is something of a godsend. The more variability in research operations, presumably the more variability in processes reported. To the extent that the operations in the various studies have sampled well the entire range of operations possible in such studies, the more likely the processes observed will be representative of the entire range of conscious processes during reading.

Thus, Peter Afflerbach and I located every verbal protocol study of reading that we could (Pressley & Afflerbach, 1995). Our intent was to categorize and organize every conscious reading process reported by the readers in these studies. That there were 40 studies made this task formidable. These studies were produced in diverse disciplines, from cognitive psychology to rhetoric to reading education and thus, ranged from extremely quantitative efforts to entirely qualitative investigations. That there were very different reporting standards and practices across studies greatly increased the challenge in summarizing data across investigations.

Even so, Pressley proceeded to do so, using a constant comparison approach, checked and challenged by Afflerbach. Pressley and
Afflerbach (1995) read every study completely, initially listing every process reported in the studies. These lists were then reduced by collapsing over redundant reports, categorized, and then organized. The categorizing and organizing continued until neither Pressley nor Afflerbach could discern any new categories or relationships between categories.

The final result required 27 single-spaced pages typed in a small font (see Chapter 3, Pressley & Afflerbach, 1995). Since producing that catalog of conscious reading processes, another half dozen verbal protocol studies have come to my attention. None of those included data that would have altered the 27-page summary. Thus, our confidence is increasing that we exhaustively categorized conscious reading processes. Table 2 is a much reduced version of the full Pressley and Afflerbach (1995) catalog.

Table 2

I. Identifying and Learning Text Content

A. Before Reading

1. Constructing a goal for reading of this text
2. Overviewing (skimming) the text
3. Deciding to read only particular sections and which particular sections
4. Deciding to quit the reading because content irrelevant to reading goals
5. Activating prior knowledge and related knowledge
6. Summarizing what was gained from previewing
7. Based on overviewing, generating an hypothesis about text meaning

B. During Initial Front-to-Back Reading

1. Generally front-to-back (i.e., linear) reading of text
2. Reading only some sections, ones believed to contain critical information
3. Skimming (i.e., less complete than front-to-back skimming cited earlier)
4. If text is easy, reading using automatic processes, until something goes wrong
5. Reading aloud; voicing what is otherwise subvocal speech
6. Repeating/restating text just read to hold in working memory
7. Repeating/restating a thought that occurred during reading

continued....
Table 2 (continued)

8. Making notes
9. Pausing to reflect on text (and perhaps notes, if made)
10. Paraphrasing part of text
11. Explicitly looking for related words, concepts, or ideas in text and using them to construct a main idea, gist, or summary
12. Looking for patterns in the text
13. Predicting-substantiating (i.e., draft-and-revision strategy for main ideas of text as well as how the author has structured the text)
14. Resetting reading/learning goals at a different level of understanding because the text suggested that there might be a more appropriate goal

C. Processes in Identifying Important Information in Text

1. Looking for information relevant to personal or professional goals or specific reading goals for this text (i.e., reading selectively)
2. Deciding which pieces of information in text are important (in relation to the goal involved in reading this text), based on prior knowledge
3. Looking specifically for what is “news” in the reading
4. Dismissing information presented in text because it is not consistent with prior knowledge (i.e., accepted thinking in domain covered by the reading)
5. Looking for/acquiring key words (i.e., concepts repeated in text; important vocabulary, phrases; qualifying words, such as if, when, only)
6. Looking for topic sentences
7. Looking for topic paragraphs
8. Noting parts of text to remember for future reference
9. Noting references in the text that should be looked at or considered later
10. Somehow marking important points in text, including important examples
11. Skipping examples because general points not provided in examples
12. Copying key sentences
13. Adjusting importance ratings as additional text is encountered continued....
Table 2 (continued)

D. Conscious Inference-Making
1. Inferring the referent of a pronoun
2. Filling in deleted information
3. Inferring the meanings of words based on clues
4. Inferring the connotations of words and sentences in the text
5. Relating information encountered in text to prior knowledge
6. Making inferences about the author
7. Making inferences about the state of actors, world in a text
8. Confirming/disconfirming an inference with information in subsequent text
9. Stating/drawing of/deducing implied conclusion

E. Integrating Different Parts of Text
1. Explicitly attempting to get the "big picture" before worrying about details
2. Generating the big idea as well as the development of ideas about component parts, with these related to one another during the reading of the text
3. Noting different parts of text and their inter-relationships
4. Holding representations of the ideas developed in text in working memory
5. Combining text structure and contextual clues to determine text meaning
6. Searching text for information related to point currently encountered
7. Searching text after a first reading, hoping to find/stimulate a macrostructure, because a satisfactory one was not detected during first reading
8. Rereading text to search for intersentential connections
9. Relating the currently read text to a previous portion of text
10. Making notes to assist/stimulate integration

F. Interpreting
1. Paraphrasing parts of text into more familiar terms
2. Visualizing concepts, relations, emotions specified in/inferred from text
3. Identifying "symbols" or "symbolic language" and translating them

continued....
Table 2 (continued)

4. Instantiating prior knowledge schemata that are activated by information in the text (e.g., thinking about a particular restaurant while reading an article about the social hierarchies in restaurants)

5. Empathizing with messages in text

6. Making claim about “what the author really wanted to say”

7. Constructing interpretive conclusions

8. Constructing interpretive categorizations (e.g., of the entire text type; of general concepts developed in text)

9. Physically or mentally doing (enacting) what the text instructs the reader to do (or suggests people should do) and then confirming the expected outcome or noting the discrepancy from the expected

10. Constructing (and/or holding in memory) alternative interpretations of text

11. Constructing (and/or holding in memory) alternative perspectives on a story from the perspectives of different characters in the tale

12. Pretending to deliberate with others while reading the text, perhaps by talking to themselves, with alternative interpretations entering the dialogue

G. After Reading

1. Rereading after the first reading

2. Reciting of text to increase memory of it

3. Listing pieces of information in text

4. Constructing cohesive summary of the text

5. Self-questioning, self-testing over text content

6. Imagining how hypothetical situations might be viewed

7. Reflecting on information in article, with it possible for consequent shifts in interpretation unfolding over an extended period of time

8. Rereading parts of text following reflection

9. Continually evaluating and possibly reconstructing understanding

10. Changing one’s response to a text as the understanding is reconstructed

11. Reflecting on/mentally recoding text in anticipation of using it later

continued....
II. Monitoring

A. Text Characteristics: Perception of...
   1. Whether text content is relevant to the reading goal
   2. Difficulty of the text
   3. Author's style/style of text; structure of the text
   4. Linguistic characteristics of text (e.g., lexical-morphological, syntactic)
   5. Specific biases reflected in text content, specific expectations of the text author about the readership
   6. Relation of this part of text to larger themes in the text
   7. Relation of this text to other sources
   8. When text is ambiguous or potentially so
   9. Relationship between own background knowledge and text content
   10. Tone of the text

B. Meaningful Processing of Text: Perception of...
   1. One's purpose in reading the text
   2. Own behaviors/strategies in processing text
   3. Reading behaviors/strategies as in the service of the reading goal
   4. One's typical reactions to this type of text
   5. The difference in reaction to this text compared to typical reactions to text
   6. Effectiveness of processes and strategies used to make meaning
   7. When comprehension processes are challenging capacity limit
   8. When there has been progress in meaning-making, although more to go
   9. Whether overall meaning is comprehended or reading goal is accomplished
   10. Text gets easier to read as meaning becomes more certain
   11. When the end of a unit of meaning has occurred
   12. When the reading goal has been achieved

C. Problems: Recognizing...
   1. Loss of concentration
   2. Reading too quickly (e.g., decoding is occurring, but comprehension is low)

continued....
3. Text is poorly written  
4. Unfamiliar terms in text  
5. Failure to understand what has been read or achieve one’s reading goal  
6. Lack of background knowledge is affecting comprehension negatively  
7. Inconsistency between personal beliefs and information in text; inconsistency between text meaning and opinions of authoritative sources  
8. Inconsistency of one’s expectations about meaning and information in text; conflict between interpretation made previously and new information in text

D. Monitoring and the Stimulation of Cognitive Processing: Activation of Processes to Accommodate Text Characteristics/Task Demands

1. Subjects make decision about how much to interpret text strictly or liberally, depending on their goal in reading or task demand that is on them  
2. Decision to rank order reading tasks or goals based on judgment that not all are attainable or doable given contextual constraints  
3. Decision to skip material  
4. Decision to skim material  
5. Decision to read material carefully  
6. Decision to construct the meaning of text carefully because aware that the text is difficult (e.g., abstract, torturous syntax)  
7. Decision to reset reading goal at a lower level because it is apparent that the reader will not be able to fulfill original reading goal by reading this text  
8. Decision to look up background material in other sources because aware that other knowledge is required to make sense of what is in a current text  
9. Decision to dispense with processing of some part of text because of awareness of potential capacity overload  
10. Decision to focus on some content and not other material because of beliefs about processing strengths and weaknesses

continued....
Table 2 (continued)

11. Decision to reread material in one section because it is not yet understood
12. Decision to reread material in one section because it is interesting
13. Decision to just keep reading in hope that later content will become clearer
14. Attempt to pinpoint confusions

E. Activation of Processing Due to Awareness of Difficulties at Word or Phrase Level

1. Evaluating importance of unknown word or phrase to overall meaning of text before deciding whether to expend effort to determine its meaning
2. Greater attention paid to unknown word or phrase
3. Use of context clues to interpret a word or phrase
4. A candidate meaning for unknown word/phrase is generated, with subsequent evaluation of the reasonableness of the sentence using that meaning
5. Generating hypotheses about confusing word, concept, or phrase followed by attempts to determine the adequacy of the hypothesis
6. Just keep reading, forgetting about the word
7. Use a dictionary

F. Activation of Processing Due to Awareness of Difficulties in Understanding Meaning Beyond the Word or Phrase Level

1. Although aware of the comprehension difficulty, doing nothing
2. Once aware of a comprehension difficulty, taking a corrective approach (e.g., analyzing carefully information read thus far; rereading last section read)
3. Once several potential interpretations of text are recognized, ones not obviously consistent with one another, reader responds (e.g., constructing inferences to account for the perceived discrepancies)
4. If a part of text cannot be understood completely, shifting focus to other parts of the text or questions not considered but also need to be resolved
5. If a text cannot be understood, attempting to think of an analogy

continued....
6. If a reading-related goal is determined unattainable, adjusting the goal
7. Looking up some of the references cited in the write-up or seeking other information from other sources
8. Reading on without figuring out interpretation when one cannot be discerned
9. Distorting some information to interpret consistent with tentative hypothesis
10. Distraction (thinking about things other than reading; falling asleep)
11. Simply giving up on understanding the text and quitting

G. Post-Reading Monitoring and Decisions to Process Additionally
1. If reader is aware that the macrostructure active at the end of reading is consistent with text, and important questions that came up during the reading have been answered, not likely to search text additionally
2. If reader senses inconsistency between macrostructure active at end of a reading and text, or important questions that came up during reading have not been answered, reader continues search for meaning

III. Evaluating
A. Consistent Evaluative Mindsets
1. Anticipatory evaluation/affect, based on feelings about/knowledge of topic
2. Acceptance
3. Skepticism, with wariness heightened to the extent that the material is likely to impact conclusions considered important by the reader
4. Reader acutely aware document was written by a particular person with particular biases, purposes, background knowledge and hence, believes document must be evaluated by implied meanings

B. Focussed Evaluations
1. Style of the text
2. Content of the text
The Nature of Conscious Reading Emerging from the Verbal Protocol Data

What is most striking from reading the verbal protocol studies is that readers are so driven to construct meaning. Every action in Tables 1 and 2 are directed at meaning making or is the result of meaning making. Readers can interact flexibly with text, using their prior knowledge to construct interpretations of what they are reading, relating what they already know to the new ideas in text. Readers often respond passionately to ideas conveyed through text. As Pressley and Afflerbach (1995) noted: They are constructively responsive in their reading, especially when they are working with texts that are important to them, interesting to them, and related to matters in which they have decidedly well-informed opinions and clear expertise.

There are four clear indications of constructive responsivity in the verbal protocols:

1. Readers can actively search, reflect on, and respond to text in pursuit of main ideas and important details. The skilled reader comes to a text knowing that it has main ideas and supporting information. An overview of the text can provide a great deal of information about the general type of information covered in it and where various topics in the reading are located. There is definitely differential attention to information in text that seems centrally relevant to the reader’s goals. The reader sometimes jumps back and forth to consider important points in the text carefully, points that seem critical to comprehend in order to get what seems like critical information from text. There is focus on important details as part of constructing the whole meaning of text.

Inferential activities also reflect the pursuit of larger themes, from inferences about the author’s overall intent in writing the piece to the drawing of conclusions strongly implied by the text. Readers’ awareness that the parts of text add up to a much greater whole is reflected by their many attempts to integrate across disparate parts of readings. Those attempts also reflect determination to get at the larger meaning of text, for protocol study participants reported great efforts expended in comparing parts of text, holding disparate ideas in working memory while searching for related ideas throughout text, and rereading to clarify how previously encountered information related to parts of text just covered.

After a text has been read, additional reflection and rereading are common, again in the service of finding the larger meanings in the text. Readers monitor whether they have comprehended a reading. If
they feel they have not comprehended the text’s overall meaning, this
can be motivation to process the text additionally and/or differently
in order to construct a more complete understanding of it. Evalu­
ations of the whole text are common in reader remarks, including
evaluations of the validity, interestingness, structural integrity, and
sophistication of the overall text. In short, there is construction and
response throughout the process of reading for understanding, with
pursuit of an understanding of the whole stimulating much process­
ing and analysis of the parts of text.

2. Readers respond to text with predictions and hypotheses that
reflect their prior knowledge. This can start with an overview of text,
with hypotheses advanced about the potential meaning of the text.
Hypothesis generation continues as front-to-back reading begins. At
some point, information will be encountered making clear that at least
some of the hypothesized points are in error. Is there anything
dysfunctional about prediction errors? No, they reflect active engage­
ment, attempts to understand text by relating it to prior knowledge.
That such errors were common in the think-alouds makes clear the
constructive nature of the reading captured in the protocol analyses.
That the initial hypotheses of the readers did not prevail but yielded
to information in the text makes clear the responsive nature of
consciously controlled reading.

3. Readers often are passionate in their responses to text. It
particularly comes through when readers have great expertise related
to and interest in the topic of the text. There was surprise, laughter,
puzzlement, frustration, and anxiety in the think-aloud reports. These
responses were possible because of extensive prior knowledge and
the related values and beliefs of the readers in these studies. The
passions are responsive in that they were elicited by particular points
made in text.

4. Readers’ prior knowledge predicts their comprehension pro­
cessing and responses to text. Thus, the initial hypotheses about the
meaning of text that result from overviewing are a product of associ­
ative responses to information encountered during the preview. As
reading proceeds, additional associative responding based on prior
knowledge is common. Also, prior knowledge affects decisions about
what is potentially important (e.g., novel) in a text and worthy of
differential attention and what is not so worthy. Such inferences are
largely based on prior knowledge. For example, conjectures about
Michener’s purpose in writing his current book are informed by
knowledge of Michener’s purposes in writing previous books. Inter­
pretive categorizations of a work (e.g., a “political satire” or an
“historical fiction”) require knowledge of such genres. In fact, interpretations of all sorts require prior knowledge that permits the reader to imagine the state of affairs depicted in the text as well as how the state depicted in the text contrasts with other states of affairs. Thus, it is impossible to come to an interpretation of the importance of the Kennedy presidency without knowledge of other presidencies.

Comprehension monitoring is largely enabled by prior knowledge. Much of deciding whether text is comprehended is based on whether the message abstracted from the text makes sense relative to what the reader already knows about the topic of the text. Monitoring also involves awareness of how the new information relates to old knowledge and whether one’s personal prior knowledge permits full appreciation of the text.

Evaluative responses to a text are not possible without massive prior knowledge. Judgments about the qualities of a text depend on knowing a great deal about how texts can be (and typically are) written and about previously existing ideas relevant to the text. Readers embrace pieces that are consistent with what they believe already and often reject writing that is filled with information inconsistent with their own views of the world.

Concluding Comments

Pressley and Afflerbach (1995) offered the most comprehensive analyses of the conscious processes of reading ever compiled. In doing so, they subsumed a number of other theories about comprehension, making clear that support for each could be found in the verbal protocol data but that none of them were sufficient to explain the complicated articulation of processes documented in the think-aloud studies: These included reader response theory, metacognitive theory, schema theory, propositionally based theories of discourse and inferential comprehension, and sociocultural models of comprehension (see Pressley & Afflerbach, 1995, chapter 4).

How good is the theory of constructively responsive reading proposed by Pressley and Afflerbach? It is very credible. That verbal protocols generated since the Pressley and Afflerbach (1995) model was completed seem to be consistent with reading as depicted in Table 2 increases confidence in the model. Because the data informing the model came from studies that varied so much in their particulars suggests that the model summarized in Table 2 is transferable. All of the main categories in Pressley and Afflerbach (1995)—such as the main categories summarized in Table 2—were supported by indica-
tions from multiple studies, indications that varied because of the operational variability across studies and thus, we believe the model has confirmability. That is, its various claims have been triangulated. This is a powerful model, with its power largely because it is a theory that was completely derived from data, with the grounded theorists (Pressley and Afflerbach) reflecting on, categorizing and recategorizing, and organizing and reorganizing until there was a framework that convincingly included all the data. It is a theory worthy of additional testing—which is the product grounded theory analyses are intended to produce.

IMPLICATIONS

What is different because of the analyses reported here? Quite a bit, with this work having implications for the conduct and analyses of future verbal protocol studies of reading, development of standardized measures of reading comprehension, and meta-analytic studies of complex cognitions and behaviors.

Verbal Protocol Studies of Reading

One of the most disappointing aspects of the many verbal protocol studies of reading is that the research was not very analytical. Typically, there was only one condition in a study, that is, no experimental manipulations that would permit assessment of the determinants of comprehension processing. There are very good theoretical and pragmatic reasons to believe that reading will vary as a function of reader characteristics, for example, readers’ purpose, prior knowledge, state (e.g., alert vs. tired), and motivation. Comprehension processing probably also varies as a function of external demands, such as the amount of time available for reading.

Other environmental variables may make a difference, too, such as whether text is presented linearly, as it is on a computer screen, or in a traditional book. In short, there is much to be understood about how reading varies as a function of a variety of variables.

One reason researchers have not conducted more analytical studies in the past has been that scoring verbal protocol data has been a major hassle. Particularly relevant here, every new investigation involved a great deal of effort to design an effective scoring scheme. I believe that the existence of the Pressley and Afflerbach (1995) catalog of conscious reading processes will make it much easier to conduct verbal protocol studies because it makes so clear just what processes are possible.
Although I am not optimistic that it will prove easy to design reliable scoring schemes based on the most fine-grained categorizations in Pressley and Afflerbach, I do think that reliable classifications will be possible at more coarse levels of analysis—for example, perhaps about as coarse as the categorizations offered in Table 2. Why do I think that? Examine the level of analysis in Wyatt et al. (1993) summarized in Table 1. The categorizations in that study were very reliable. The difference made by Pressley and Afflerbach (1995) is that the painstaking efforts in Wyatt et al. (1993) to develop a scoring scheme should be less painstaking in the future, so that research resources can be redirected, for example, to expand the number of conditions in these studies.

If I were doing a verbal protocol study today, and I expect I may be back doing them in the near future, I would take the Pressley and Afflerbach (1995) catalog and begin to score my protocols. Then, I would collapse over subcategories until my scoring scheme was reliable. This process should be much easier than starting from scratch and building a scoring scheme. Frankly, I cannot wait to have an opportunity to do this.

Standardized Measures of Comprehension

Standardized measures of comprehension typically require readers to read text and then answer comprehension questions. The quality of comprehension processing is then inferred from performance on the comprehension questions. In contrast, the measurement community increasingly embraces more authentic approaches to assessment, including performance assessments aimed at elucidating more directly cognitive processes. A major challenge to the development of such assessments is the scoring of them.

Just as the Pressley and Afflerbach (1995) catalog should serve researchers collecting verbal protocols, it should also make it easier to score verbal protocols that are collected as part of efforts to develop more authentic comprehension assessments. To be certain, there are enormous challenges that remain for such assessments to become a reality, but I am confident that the effort will lead to important insights about reading. For example, when Cordon and Day (1995) asked college students to think aloud as they read passages from standardized comprehension measures, processing proved to be much less sophisticated than the type of reading Wyatt et al. (1993) observed: The college students relied heavily on rereading rather than more active, selective processing.
One of the main reasons I believe that the Pressley and Afflerbach (1995) catalog will have an impact on standardized testing is that it is the test construction community who have talked most with Peter and me about *Verbal Protocols of Reading*. I am heartened by this development, and by the invitation of the Buros Institute to present here, because I am convinced that cognitive psychology has much more to offer the assessment community than it has offered in the past. In particular, my reading of the interest in *Verbal Protocols* by measurement professionals is that the Pressley and Afflerbach (1995) analysis makes obvious the possibility of reasonably easy and reliable scoring of verbal protocols of reading: One obvious possibility would be simply to collapse across the many categories of response summarized in the book until scoring was reliable.

Consider the following examples based on the level of detail in Table 2. Even if scorers had difficulty determining whether an inference involved inferring the meaning of a word based on context clues versus inferring the connotation of a word, they very likely would have no trouble agreeing that the inference was a conscious inference. Even if scorers could not agree that a particular response reflected the reader making a claim about "what the author really wanted to say" versus the reader empathizing with messages in a text, the scorers would likely agree that the reader was interpreting the text.

In summary, the verbal protocol approach offers a much more direct window on processing than other forms of comprehension measurement. At a minimum, because cognitive psychologists have collected verbal protocols of excellent reading, composing, and problem solving, those devising tests to evaluate the sophistication of cognition at least have a better understanding of the nature of sophisticated cognition than they did before the verbal protocol approach was employed extensively by cognitive psychologists. Moreover, when verbal protocols are used to assess processing, the assessment is much more driven by what is in the head of the reader, writer, or problem solver than by what in the head of the individual constructing the assessment instrument. For example, a multiple-choice item to assess whether inferencing is occurring during reading includes one logical inference, based on the item writer's perspective. The inferences an item writer makes, however, are not always the ones any given reader makes. Some readers may fail an item tapping inferencing not because they are not reading actively and making inferences but because they are not making the inferences the test constructor made. In short, I believe that verbal protocols may permit assessments of processing that are much more realistic than the assessments of the past.
Meta-Analytic Summaries of Cognitive Processes and Behaviors

There has been an explosion of interest in the past 15–20 years to find ways to summarize findings from a large number of studies. Much progress has been made in the development of meta-analytic procedures for quantitative data. In contrast, there has been little progress in finding ways to summarize qualitative outcomes.

I believe that many of the analytic procedures (Miles & Huberman, 1994) that can be used to organize qualitative data in individual studies can also be applied across studies. Pressley and Afflerbach (1995) used one approach, grounded theory analyses based on the method of constant comparison (Strauss & Corbin, 1990), to organize the data in verbal protocols of reading. I am looking forward to much more complete grounded theories of problem solving and composition processes in the near future, for the fairly large think-aloud literatures on problem solving and composition are now being analyzed by others using procedures similar to those Afflerbach and I employed. I hope this is the start of a trend.

I also think that the methods used by Pressley and Afflerbach (1995) will prove more broadly applicable, perhaps useful whenever complex behaviors are reported categorically across a number of investigations. With the expansion of qualitative methods in general, many complex processes and behaviors will be studied qualitatively, with a number of problems studied in a large number of studies. It is essential to find ways to summarize the data collected in these efforts and thus, Pressley and Afflerbach (1995) is probably the first in a long line of qualitative meta-analytical investigations.

Concluding Comment

My colleagues and I have never been afraid to break with traditional methodologies if there was promise of conceptual advance. In general, although others have embraced our findings, they typically have not followed our methodological leads. One reason is that my associates and I have never shied away from labor-intensive methods, ones requiring more data collection and more intense data collection than often occurs in social sciences and educational research—for example, my work on monitoring of strategy efficacy (e.g., Pressley & Ghatala, 1990; Pressley, Levin, & Ghatala, 1984) and the research on transactional strategies instruction (e.g., Pressley et al., 1992). Another is that I never really saw myself as a methodologist and thus, did not attempt to impress my methods on others. This time, the analyses are as labor intensive as ever, but I am more determined to
persuade others of the power of the methods my associates and I are using. I believe that there will be real advances in understanding of complex cognitive and behavioral processes if others do follow the leads of Wyatt et al. (1993) and Pressley and Afflerbach (1995). A great deal was learned about conscious text processing in the work summarized in this chapter. Moreover, the efforts to do meta-analyses now will pay off in less diagnostic effort in future research as well as more valid standardized assessments. The promise is great for empirical and theoretical advances as well as for practice.

REFERENCES


Pressley, M., Ghatala, E. S., Woloshyn, V., & Pirie, J. (1990a). Being really, really certain you know the main idea doesn’t mean you do. *Yearbook of the National Reading Conference, 39*, 249-256.


This chapter attempts to consolidate the diverse opinions and conclusions included in the previous six chapters of this volume. I have found it easiest to do so in three sections. Section 1 provides a summary the book’s main themes. These themes pertain to the need for a more comprehensive theory of metacognition, the disparity between metacognitive theory and measurement, methodological questions about the measurement of metacognitive processes, concerns about poor instrumentation, the generality of the metacognition construct, and issues pertaining to educational practice. Section 2 raises concerns central to the measurement community in general. These concerns include questions about the reliability and validity of assessment techniques and paper-and-pencil measures. Another concern is the need for dependable performance assessment of metacognitive skills among younger and older students. Section 3 makes a number of suggestions for future research and measurement practice based on current theory. A number of educational implications are discussed as well.

Six Emergent Themes

Most researchers studying metacognition agree that it is important to study, but difficult to measure. The chapters included in this volume provide a variety of strategies to bridge the gap between
metacognitive theory and measurement practice. I believe there are six discernable themes that emerge from these chapters that, collectively, point the field in a sensible direction for further research and discussion.

Theme 1: The need for a comprehensive, unified theory of metacognition. Throughout this volume, there are repeated references to two compatible, yet theoretically distinct, theories of metacognition associated with the work of John Flavell (1987) and Ann Brown (1987). Most experts, including those represented in this volume, view this as a serious problem because it prevents researchers from agreeing upon basic metacognitive processes and terminology for those processes. The terminology problem is especially important because many researchers describe the same basic processes using somewhat different terms.

However, all the contributors to this volume, as well as other theorists not included (Chi, 1987; Garner, 1987, 1994; Nelson & Narens, 1994; Metcalfe, 1994a), agree on the primacy of three overarching processes I refer to as regulatory control, performance monitoring, and task monitoring. The former refers to a variety of self-regulatory processes used to actively, and often intentionally, control cognitive activity. Many of these activities have been described in detail in related volumes on self-regulated performance (Schunk & Zimmerman, 1994; Zimmerman & Schunk, 1989) and reading (Pressley & Afflerbach, 1995; Pressley, Harris, & Guthrie, 1992). Performance monitoring refers to monitoring ongoing comprehension via either self-report or various subjective measures (see Schraw, Wise, & Roos, this volume). Task monitoring refers to assessing the demands of the task at hand, especially its difficulty relative to one’s own skills and knowledge (see Borkowski, Chan, & Muthukrishna, this volume).

A great deal of discussion centers around these processes either directly or indirectly. Most chapter authors believe the three processes are related in a reciprocal, interactive fashion, forming a triarchic set of basic metacognitive processes (Pintrich, Wolters, & Baxter, this volume). There are several explicit mentions of the relationship between control and performance monitoring, a relationship Schraw et al. (this volume) refer to as a regulatory loop. Borkowski et al. (this volume) also address the crucial relationship between task monitoring and control, especially among younger students and older students who find academic work quite difficult (Borkowski & Thorpe, 1994).

A closely related problem is the fuzzy boundary that separates overlapping constructs such as metacognition, executive processes, and self-regulation (Pintrich et al., this volume; Zimmerman, 1994).
Although little is said about this issue directly, there is some discussion regarding the relationship between metacognition and motivational variables. These authors generally agree that motivation is an important contributor to metacognition, although it is less clear whether it is part of metacognition per se (see Garcia & Pintrich, 1994 and Pressley, Borkowski, & Schneider, 1987 for related discussions).

Theme 2: There is a large discrepancy between metacognitive theory and measurement practice. Chapter authors unanimously agreed that metacognition is an important theoretical construct that warrants serious attention. They also agreed that measurement aspects of metacognition typically are included as an afterthought. Most researchers are interested in measuring metacognition only as far as those measurements can be used to evaluate metacognitive theory or improve instructional practice. This perspective limits focus on the integrity of the measurement process apart from how it impacts metacognitive theory.

Reducing the gap between metacognitive theory and measurement practice depends in part on the success of establishing a comprehensive theory of metacognition that researchers can use as a common referent point (see Theme 1). It also depends on researchers’ willingness to become conversant with contemporary measurement theory and to value the integrity of one’s measurement instruments as much as one’s metacognitive theory. This is a tall order at the present, given that many researchers interested in metacognition have limited training in psychometric theory, little interest per se in measurement, and tend to view measurement as an unglamorous means to a more impressive theoretical end.

Theme 3: There is considerable debate regarding the relative pros and cons of different assessment methodologies. As in any contemporary debate on research, this volume present differing views regarding the utility of qualitative and quantitative methods. Pressley (this volume) presents a careful argument for the increased use of qualitative methods such as grounded theory approaches Strauss & Corbin, 1990). According to Pressley (this volume), researchers have privileged access to the thoughts and strategies used by individuals as they perform a task or reflect on their performance.

Others were less positive about qualitative approaches, however voicing traditional concerns summarized in Ericsson and Simon (1993) and Pressley and Afflerbach (1995). These include the potentially intrusive nature of interviews, the possibility that individuals do not have privileged access to their own cognitive processes, and the possibility that individuals will provide biased reports of their activi-
ties. Nevertheless, all of the contributors to this volume agreed that both traditions have strengths and weaknesses, and do not guarantee informative or accurate assessment. More importantly, there was strong agreement that any method is superior to no method at all! The most reasonable strategy at this point is to recognize the problem, then turn our attention to the daunting task of unifying metacognitive theory and constructing reliable and valid methods for evaluating that theory.

Another methodological subtheme concerned what several authors referred to as grain size; that is, the specificity of the task that is being assessed (Howard-Rose & Winne, 1993; Pintrich et al., this volume). Pressley (this volume), for example, described ongoing research of an extremely broad grain size—no less than an exhaustive taxonomy of strategies used during highly constructive reading. Pressley and Afflerbach (1995) reported over 150 separate strategies. Tobias and Everson (this volume) reported studies of much smaller grain size; that is, the accuracy of specific monitoring judgments. The process model described by Borkowski et al. (this volume) necessitated a number of assessments, each of differing grain sizes.

The issue of grain size is important for two reasons. One is that bigger grains become increasingly dependent on longitudinal, qualitative methods, although it is possible to use sophisticated quantitative methods such as structural equation modelling as well. In contrast, very small grains seem easier to study with traditional quantitative methods such as reaction and choice selection times, recognition and recall measures, or calibrated accuracy judgments. One potential drawback, however, is that small-grain processes may be so automatic, individuals may no longer have privileged access to performance, precluding the use of introspective measures.

Grain size also affects the degree to which measurements provide a useful test of metacognitive theory or educational interventions. For example, although studying the acquisition of a specific strategy is crucial (see Borkowski et al., this volume), such information would not be sufficient to assess the validity of the process models described by Pressley and Afflerbach (1995) or Borkowski et al. (this volume).

A third methodological subtheme concerned what kind of metacognitive processes individuals choose to study. The general consensus was that small-grain processes (e.g., monitoring strategies) vary among different groups and ages, and that one model of metacognition may not apply to all individuals. On the other hand, studying monitoring processes in experts and novices may lead to different results (Glaser & Chi, 1988; Glenberg & Epstein, 1987). The
current volume does not offer a resolution to this paradox. I believe that researchers will have to make a greater effort to construct a comprehensive developmental model of metacognition as opposed to a purely descriptive account of what the average college sophomore does. Clearly, models are needed in which expert and novice performance can be reconciled.

At this point, there continue to be important differences of opinion regarding how to measure metacognition. This issue must be resolved before substantial progress can be made in the field. Specifically, researchers must at least agree on what constitutes necessary and sufficient evidence for assessing the validity of metacognitive constructs, even if researchers continue to disagree about the methods they use to collect evidence. Pintrich, Wolters, and Baxter (this volume) make a number of important suggestions in this regard based on the work of Messick (1989).

Theme 4: Most available instruments that measure metacognition have unknown psychometric properties. Both Pintrich et al. (this volume) and Baker and Cerro (this volume) suggested that most measures of metacognition can be characterized by two salient features: (a) they were constructed for use within a specific study and (b) there is little or no normative information about them even within the population for which they were designed. There are two measurement-related consequences. One is that the dimensionality of the instruments (i.e., what psychometricians would refer to as an instrument’s factor structure) is unknown. This prevents researchers from identifying the number and type of psychological constructs the instrument presumably measures. Second, there is no information regarding how the hypothesized construct is related to other relevant performance outcomes.

There are several instruments that prove exceptions to this rule. One is the Learning and Study Strategies Inventory (LASSI) developed by Weinstein, Zimmerman, and Palmer (1988), which reports 10 separate subscales, including attitude, motivation, anxiety, test strategies, and self-testing. Although the LASSI has acceptable internal consistency measures for each scale (i.e., $\alpha = .70$ to .80), and correlates with measures of cognitive performance, it is unclear whether it measures metacognition per se, or cognitive skills such as study strategies that are regulated with the help of metacognitive knowledge. Another instrument is the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich & DeGroot, 1990; Pintrich, Smith, Garcia, & McKeachie, 1993), which measures both motivational and strategy subscales. A third scale is the Metacognitive Assessment Inventory (MAI) (Schraw & Dennison, 1994), which includes knowl-
edge of cognition and regulation of cognition subscales. Although the MAI has extremely high reliability, its relationship to other cognitive measures remains to be investigated.

As this list indicates, all of these instruments have been developed within the last decade, and most are fewer than five years old. Initial results seem promising, and there is complete agreement that all of these instruments represent positive steps in the assessment of basic metacognitive knowledge. However, as Pressley and Afflerbach (1995) point out, paper-and-pencil inventories are no substitute for more in-depth analyses of metacognitive knowledge. Second, inventories alone seem incapable of capturing the complex dynamics of metacognitive regulation. Third, all of these instruments are intended for use with adolescents and adults; thus, instrumentation is needed for younger learners.

Theme 5: There is uncertainty (and discomfort) regarding the domain-generality of metacognition. The degree to which metacognition constitutes a domain-general phenomenon remains an important question. The basic issue is this: If metacognition is really a function of expertise, and therefore is domain-specific, can it be measured at all using a domain-general instrument such as the LASSI, MSLQ, or MAI? An equally pressing problem is how to test whether metacognition is a domain-general phenomenon? Several studies have addressed this question either directly or indirectly. Glenberg and Epstein (1987), for instance, found a negative relationship between expertise and monitoring. Music experts monitored more poorly than physics experts on a test of music principles whereas the reverse was true on a physics test! Morris (1990) found that domain knowledge was unrelated to monitoring proficiency even though it was related to one’s ability to answer questions effectively in that domain. However, Maki and Serra (1992) found that monitoring improved as individuals acquired more information from the to-be-learned materials. These studies provided indirect, as well as conflicting, evidence about the domain-specific nature of metacognitive knowledge and regulation.

A more direct study was conducted by Schraw, Dunkle, Bendixen, and Roedel (1995). In the first of two experiments, individuals completed eight multiple-choice tests that varied with respect to content domain, overall difficulty, number of items, inferential difficulty, and number of distractors. Despite the heterogeneity of these tests, confidence judgments were correlated in the neighborhood of $r = .50$ among all tests even when performance was controlled statistically. Monitoring accuracy and discrimination scores (i.e., the ability to discriminate between correct and incorrect answers) were also
correlated. However, in Experiment 2, where all five tests were matched on length, difficulty, and test-item format, monitoring accuracy and discrimination scores were correlated even when performance was controlled. Schraw et al. (1995) concluded that monitoring accuracy and discrimination were attributable to two processes that included domain-specific expertise and knowledge that supports performance, and domain-general regulatory strategies such as self-checking that supports self-regulation.

Of course, all of the studies described in this section investigated metacognitive processes of limited grain size (i.e., comprehension monitoring). Thus, even though there is evidence that comprehension monitoring relies in part on domain-general metacognitive knowledge, it does not follow that other metacognitive processes (e.g., planning, and allocation of cognitive resources) are domain-general.

The theoretical and educational implications of future research in this area are crucial in my opinion. One implication of data consistent with the domain-general view is that educators may feel more confident teaching domain-general metacognitive skills, rather than skills that are encapsulated within specific domains such as mathematics and reading (see Fodor, 1983; Gardner, 1983; and Hirschfeld & Gelman, 1994, for opposing views). A second implication is that researchers must inquire about the development of domain-general skills. One intriguing explanation has been proposed by Karmiloff-Smith (1992) who hypothesized that domain-specific skills and knowledge are merged over time across domains to create domain-general knowledge. Understanding the underlying cognitive mechanisms that enable this development is a worthy topic in and of itself.

Theme 6: Difficulty relating metacognitive theory to educational practice. One often heard complaint is that there are too few proven methods for improving metacognition among children and adolescents. Baker and Cerro (this volume) point to a number of successes, including Palincsar and Brown (1984), Paris and colleagues (Cross & Paris, 1988; Jacobs & Paris, 1987), and Pressley and colleagues (Brown & Pressley, 1994; Pressley, Harris, & Marks, 1992). Nevertheless, Baker and Cerro also are outspoken about theorists' willingness to make snap educational recommendations without any evidence to support their claims. In fact, most training studies have reported modest, yet lasting, gains only after intensive instruction lasting from 6 weeks to 6 months (Delclos & Harrington, 1991; King, 1991, 1992).

Another major problem is the almost complete lack of standardized assessment guidelines for use in the classroom or in research settings, especially when evaluating younger learners. One notable
study in this regard was Swanson (1990), who found that metacognitive knowledge among fifth and sixth graders facilitated problem solving. Swanson also suggested that metacognitive regulation may develop independent of skills traditionally measured on aptitude tests, although this controversial claim requires additional research. To assess metacognition, Swanson administered a 20-question verbal interview, tapping student's knowledge about self-regulation. Swanson also provided extensive documentation for scoring the interview, although with the exception of Corkill and Koshida (1993), others have not capitalized on these guidelines.

SUMMARY

The six themes related to the theory and practice of metacognition provide an ambitious agenda for future theory development, research, and educational practice. It seems clear that a new agenda is in order. Indeed, it is conceivable that the construct of metacognition may lose its appeal to practitioners, and its theoretical relevance, if researchers do not provide a much greater degree of psychometric rigor during the next decade. There are several salient issues facing researchers. One is to construct a comprehensive theory that includes well-specified subcomponents. A second is to agree on what kinds of evidence are necessary to validate this theory. A third is to construct standardized procedures using qualitative and quantitative methods to assess metacognitive competencies. A fourth is to design and test interventions to improve or remediate important metacognitive competencies. A fifth is to propose standards by which these interventions can be evaluated.

Thoughts from the Measurement Community

Not everyone is excited by terms like metacognition and self-regulation. As recently as the mid-1980s, many people felt that metacognition was too broad and elusive to be studied effectively. One common complaint, and the driving force behind the symposium, is that some theorists interested in metacognition have neither the training or inclination to establish sound psychometric underpinnings to measures of the construct.

In this spirit, there are several cautions that one might expect to hear from testing and measurement experts. Although these cautions parallel the six themes described above, they provide a view of the problem from a somewhat different perspective.

Caution 1: The field needs a plan for comprehensive assessment of the construct. Test authors would quickly point to the testing industry to
make their case. Aptitude tests, for example, come in virtually every size and shape imaginable. These tests typically provide reliability and validity norms, and endeavor to meet all the standard evaluative criteria proposed by measurement theorists such as the Standards for Educational and Psychological Testing (AREA, APA, & NCME, 1985). Notwithstanding the continuing debate about the validity of aptitude tests, and the role that aptitude plays in learning (see Jensen, 1992, and Sternberg, 1986, for different views), modern testing as we know it would not be possible without an overarching plan for translating aptitude theory into instruments that can be evaluated with exacting care.

Caution 2: Generate and test models. Researchers have failed for the most part to translate metacognitive theory into testable models. In this view, a theory provides a method for systematically organizing a body of knowledge that explains a particular set of phenomena (Byrnes, 1992; Kuhn, 1989; Schraw & Moshman, 1995). Every theory has at least two distinguishable parts; a formal aspect (i.e., postulates about how a phenomenon occurs) and an empirical aspect (i.e., a test of those postulates, usually in the form of data or mathematical proofs). A model provides a formal description of a theory by specifying the relationships among its most important postulates. These descriptions often take the form of a diagram, flow chart, or summary table. Models are convenient ways to operationalize a theory and enable researchers to test one part of the theory at a time. Models are useful in that a theory can be tested and modified without discarding the entire model.

Both Flavell (1987) and Brown (1987) proposed descriptive theories of metacognition that, although overlapping, remain somewhat independent of one another. Surprisingly, very few researchers have attempted to translate these theories into operational models that enable researchers to investigate systematically the relationships among model components. Two important exceptions are the models proposed by Nelson and Narens (1990, 1994) and Borkowski and colleagues (Borkowski & Muthukrishna, 1992; Borkowski & Thorpe, 1994; Borkowski et al., this volume). These models enable researchers to make explicit predictions about the relationships among control and monitoring processes, one’s extant knowledge base, and motivational factors that affect self-regulation.

Process models clearly play an important role in theory testing. Models specify particular relationships that can be tested explicitly. Of equal importance, models focus our attention on the role that instrumentation and measurement play in theory testing. The advent and continued growth of structural equation modeling, for example,
has ushered in a new era in the measurement of salient constructs in cognitive psychology. Such models require researchers to test not only the structural aspects of the model (i.e., relationships among model components predicted by theory), but measurement aspects as well (Bollen, 1989).

Caution 3: Construct and evaluate instruments that assess specific components of the model. Three questions seem especially germane to this caution. These include: (a) What is the construct of interest? (b) Do appropriate measures exist to measure it? and (c) Does the measurement process change the construct? Each of these questions is considered briefly.

What is the construct? Broad constructs are difficult to measure, making it unlikely that there will ever be a single measure of metacognition (see Pintrich et al., this volume; Schraw, 1995). An alternative approach would be to partition the metacognition construct into smaller components, then hypothesize about the relationships among subcomponents, although others may disagree with this strategy (Linn, 1991). Constructing a detailed structural model that specifies individual components and their interrelationships is an essential precursor to the validation process.

Do appropriate measures exist? Once potentially measurable subcomponents have been defined, researchers must select behavioral and self-report indices that measure these components. Indices of metacognitive activity should be evaluated in an ongoing manner using a variety of approaches (see Crocker & Algina, 1986, chapter 4, for a summary of this process). Potential measures need to be evaluated with respect to reliability, validity, and utility, or their qualitative counterparts such as credibility and authenticity (Creswell, 1994; Merriam, 1988; Miles & Huberman, 1984).

Two aspects of reliability are essential. One concerns the internal consistency of an instrument. Measures such as coefficient alpha provide easy-to-compute indices of internal consistency and are available for a variety of measures of metacognitive knowledge and performance (Pintrich et al., 1993; Schraw & Dennison, 1994; Weinstein et al., 1988). However, when used in isolation, coefficient alpha is insufficient. A second aspect of reliability is consistency over time. Currently, there is little available information regarding test-retest reliability on commonly used measures of self-regulation and metacognition.

Multiple aspects of validity are crucial to the effective use and interpretation of scores from an instrument or experimental results. In their chapter, Pintrich et al. (this volume) addressed the question of
validity using Messick's (1989) five-component framework. In this view, all questions of validity are essentially questions related to construct validity (i.e., the degree to which inferences about a score accurately represent an observable or unobservable phenomenon of interest). Given that Messick views each of these five components as interrelated, threats to any component necessarily affect all others.

A similar argument can be made for the credibility of qualitative methods. Whether researchers use terms such as utility or confirmability is beside the point; rather, they must demonstrate that they have explained a hypothetical construct (or phenomenon) in an accurate, replicable way. Researchers with a predilection for either qualitative or quantitative methods should conduct carefully planned validation studies to assure their methods accurately describe the phenomenon of interest. In particular, there is a tremendous need for studies evaluating the convergent and divergent validity of scores from multiple measures, preferably using multiple methodologies, to explain metacognitive phenomena. Point 3 made by Pintrich et al. (this volume), in which they recommended the use of a multitrait, multimethod approach to construct validation, emphasized this concern directly.

Does the measurement process change the construct? Researchers must ask themselves to what extent their measures of metacognition affect the deployment of metacognitive knowledge and regulatory skills. Although there are a variety of potential measurement confounds, several that I consider to be especially serious are discussed below (see Baker & Cerro, this volume). The first is that self-report inventories (e.g., Schraw & Dennison, 1994) may elicit socially desirable responses. One way to safeguard against instrument bias is to conduct convergent and divergent validity studies. Unfortunately, few studies of this kind have been reported.

A second potential problem is that think-aloud studies may affect the measurement of metacognitive processes by competing for limited resources that are necessary for task performance. One option is to use retrospective self-reports; however, verbal report theorists generally view retrospective reports as less reliable than concurrent self-reports. Competition for limited resources also affects many quantitative studies, but especially those using on-line confidence judgments of performance. One alternative is to use unobtrusive measures such as computerized testing procedures (Schraw et al., this volume).

A third problem is that structured interviews may provide information to individuals that they would not report on their own, thereby masking their true metacognitive knowledge. One option is
to ask informants to respond in writing, although this leads to other problems such as competition for limited resources.

A fourth problem, and the most serious in my opinion, is that individuals may differ in terms of their ability to explain or estimate their metacognitive knowledge either verbally or in writing. For example, experts may have a richer vocabulary for describing their mental processes, even though novices engage in similar processes. In essence, this problem is due to aptitude by treatment interactions, and must be considered carefully when researchers evaluate potential threats to the validity of their findings.

Caution 4: Use diverse assessment methods. Pintrich et al. (this volume) captured the main theme of the Buros Symposium when they concluded there is no one-size-fits-all measure of metacognition. Indeed, if we accept the premise that there are two or more distinct metacognitive processes (e.g., control and monitoring), it seems reasonable to conclude that there must be different measures of these processes. Equally reasonable is the assumption that each identifiable metacognitive process can be measured using different instruments and methodologies (see Borkowski et al., this volume).

Perhaps the best advice one could give researchers interested in metacognition is to adopt a multitrait, multimethod (MTMM) model of assessment along the lines first proposed by Campbell and Fiske (1959), and elaborated upon by Cook and Campbell (1979). The MTMM approach emphasizes the collection of multiple measures of a phenomenon using multiple, preferably diverse, methodologies. In the context of validating metacognitive theory, this means using objective self-reports (Pintrich et al., this volume), subjective self-reports such as concurrent verbal reports (Pressley, this volume), subjective assessments of one’s thinking or performance such as calibration judgments (Tobias & Everson, this volume), and unobtrusive measures such as item selection times using computer-based testing (Schraw et al., this volume). Other measurement approaches are needed as well, such as neurophysiological correlates of metacognition (Metcalfe, 1994b).

Summary

The four cautions outlined above summarize basic measurement concerns for any research agenda. They have special importance for the domain of metacognitive research given the paucity of systematic validation studies. Foremost, the field needs a systematic and comprehensive assessment agenda, lest it lose its credibility among psy-
chologists and educators in general, and measurement experts in particular. A major part of this agenda should include the development of a testable model of metacognition. Work in this direction was described by Borkowski et al. (this volume). Second, working models must be tested and refined in an ongoing basis. Earlier stages of this research will undoubtedly focus on specific components of metacognition, whereas the later stages may test multicomponent models using procedures such as structural equation modeling; that is, the grain size of these studies would be expected to change over time. Third, a variety of measures should be used to assess each separate component, as well as the entire model. Last, adopting a multitrait, multimethod approach to model testing may enhance the effectiveness of this research.

An Agenda for Future Research

There are many ways that the assessment of metacognition can be improved in the future, but it is not an easy task, and it cannot be done without the help of colleagues in different fields of study. This section makes seven suggestions for future research and practice based on the cautions described above. I have rank ordered these suggestions in a way that might surprise some readers. Regardless of their ordering, all are essential to the advancement of the field, and should be taken seriously.

Suggestion 1: Researchers interested in metacognition must collaborate with measurement and instructional design experts. Most of the authors included in this volume would acknowledge their lack of technical measurement expertise. Most of them, and most researchers working in the area of metacognition, are not keenly interested in measurement issues per se. Many, including myself, do not have a strong interest in the design of instructional interventions. Yet it is clear to me that failing to involve measurement and instructional design experts may be disastrous to the field. I believe we are at an impasse that cannot be overcome without the skills and knowledge of experts who do not share our vested interest in the construct of metacognition.

On a brighter note, let me focus on some of the potential advantages of cross-disciplinary collaborations. One is that experts at solving measurement problems may add a tremendous amount of richness to existing theory. A second is that measurement and instructional design experts may provide innovative ways to assess metacognition, both with respect to objective paper-and-pencil measures, as well as performance-based assessments. A third is that
instructional design experts have much to offer by way of translating metacognitive theory into instructional practice. The same may be said regarding how to evaluate formative and summative instructional outcomes.

Suggestion 2: Agree on a unified theoretical framework. It bears repeating that the field is perceived by some outsiders as too theoretically disparate. It is unlikely that substantially more progress will be made until researchers agree on what it is they are looking for. My personal preferences are for a three-component model of metacognition that emphasizes the role of regulatory control, performance monitoring, and task-monitoring processes. Such a model is already consistent with much of the theorizing being done in metacognition and self-regulation (Baker, 1989; Borkowski et al., this volume; Garner, 1987; Garner & Alexander, 1989; Jacobs & Paris, 1987; Nelson & Narens, 1994; Pintrich & DeGroot, 1990; Pressley & Afflerbach, 1995; Schraw & Moshman, 1995). Even earlier theories such as those proposed by Brown (1987) and Flavell (1987) that postulate two main components, are highly consistent with the three-component view.

It is my view that the most important issue separating existing accounts of metacognition is terminological differences, not assumptions about basic metacognitive processes. In a way, it is as if the field has agreed on what kind of pizza it wants to eat, but still can’t decide how to slice the pieces! I believe we are much closer to a comprehensive theory of metacognition than most casual observers would give us credit for. Perhaps our greatest challenge is to recognize the perception of disarray from outside the field and to resolve it!

Another theory-related suggestion is to focus on measuring the practical and statistical relationships among the three main components described above. Recent work by Nelson and Narens (1990, 1994) and Koriat (1993, 1994) have raised extremely important questions about the relationship between monitoring and control. The importance of this relationship, as well as ways to test it, were echoed by Schraw et al. (this volume). Borkowski et al. (this volume) also emphasized the relationship between performance and task monitoring.

Suggestion 3: Identify suitable outcome measures that can be used as criteria to evaluate metacognitive behaviors. Most of the work done in the field of metacognition has been devoted to generating theory. A sizable amount of work also has been done on testing a narrow band of metacognitive processes, but especially comprehension monitoring (See Baker, 1989; Pressley & Ghatala, 1990; and Schraw & Moshman, 1995, for reviews). Some work, but not a great deal, has been done on instructional improvement (Brown & Pressley, 1994; Jacobs & Paris,
ASSESSING METACOGNITION

Yet virtually no attention has been paid to identifying suitable outcome measures of metacognition, other than as they apply to testing components of metacognitive theories, or evaluating the efficacy of instructional interventions.

Of particular concern are measures with a high degree of ecological validity, including performance-based measures that capture midrange (e.g., 3 months) and long-term (e.g., 6 months or more) development of metacognition. There are not any suitable methods in this regard at the present time, and unfortunately, few researchers seem to have considered this problem (see Baker & Cerro, this volume, and Garner, 1987, for further discussion).

Outcome measures should be identified that meet a variety of needs. One especially important need is to parallel measures of metacognition that are suitable for children and adults. Without such measures, comparisons among age groups at a single time are compromised, as are across-time comparisons within the same group. A second need is to identify measures that are suitable for field settings. Many studies rely on either checklists or overt behavioral measures. Lacking is a method for identifying multiple levels of metacognitive activity via self-report. A third need is to construct specific versus broad measures of metacognition. Most studies focus on monitoring judgments made prior to or after completing a test item. Although monitoring reflects an important component of metacognition, it fails to provide information about strategies for planning, debugging, or evaluating.

Another important question to ask about these measures is the degree to which they intrude on metacognitive processes. Asking individuals to think-aloud as they perform a task, for example, may interfere with that task by consuming limited resources, or it may facilitate metacognitive behaviors by calling explicit attention to task demands.

Suggestion 4: Empirically investigate the relationship among different outcome measures. One consequence of the small amount of work that has gone into identifying alternative outcome measures is that very little is known about the relationship among these measures. For example, there has not been a systematic comparison of the relationship between ratings of monitoring accuracy and self-reported strategies, even though these measures are used frequently in the literature. As measurement experts would remind us, the more measures we use to assess an outcome, and the more we know about the relationship among these measures, the better able we are to make reliable and
valid inferences about the construct under study. Put simply, it is crucial to converge on a construct using multiple, triangulated methods.

Suggestion 5: Establish methodological and data-analytic guidelines for measuring metacognition. Any research tradition quickly establishes its own informal or formal guidelines for administering and interpreting a task. One example is the use of monitoring accuracy judgments that pervade the metacognition literature. Nevertheless, with the exception of monitoring accuracy judgments (see Tobias & Everson, this volume), error-detection (see Baker & Cerro, this volume), and self-reported strategy use (see Pressley, this volume), there are few guidelines available for either researchers or practitioners. This is a serious obstacle for educators who want to improve metacognition in the classroom. One difficulty is understanding what metacognition is, a problem that is due in part to too much theoretical and terminological slippage. Another problem is selecting a benchmark for measuring the growth of metacognition (see Suggestion 3 above). A third problem is deciding on a method for assessing metacognitive progress. My observations have led me to conclude that teachers and parents typically assess metacognitive growth using intuitive hunches rather than tangible outcome measures. This is due to at least three factors: (a) uncertainty about what to look for; (b) the lack of meaningful, cost-efficient measurement strategies; and (c) the lack of meaningful interpretative guidelines.

The lack of procedural and interpretative guidelines is a serious problem within and beyond the field of metacognition. Within the field, there are few established procedures for measuring metacognitive knowledge. Of the three most widely used paper-and-pencil instruments (i.e., the LASSI, MSLQ, and MAI), only the LASSI has been used enough to provide relatively stable norms for different types of students. Most studies continue to use instruments designed specifically for the study at hand. Many studies that used the LASSI, MSLQ, or MAI also have fairly small sample sizes (i.e., less than 100), that preclude a meaningful replication of previous studies. To complicate matters, very few studies ever report reliability coefficients for criterion measures, and few researchers have enough measurement and statistical savvy to understand the subtle, yet important, differences among alternative data-analytic strategies once data have been collected (Keren, 1991; Schraw, 1995).

Another issue concerns the purpose of metacognitive assessment (Baker & Cerro, this volume). Information gathered during an assessment can be used for different purposes (Tindal & Marston, 1990). One purpose is to make placement decisions, such as whether a
student is admitted to a restricted program. A second purpose is to make diagnostic judgments regarding a specific skill or learning disability. A third purpose is to use the information to provide ongoing feedback to students, parents, and teachers. A fourth purpose is to provide summative evaluation at the end of an instructional unit or training session. Whereas most researchers used measures of metacognition as summative indices of unobservable metacognitive competencies, most educators are interested in metacognition from a diagnostic and formative evaluation perspective. There has been very little thought given to bridging the gap between researchers and theorists on the one hand, and practitioners on the other. I believe researchers have a responsibility to bridge this gap and to address how educators can collect information about a student’s skills in a manner that enables the educator to provide useful formative feedback. Similarly, many teachers hunger for guidelines for diagnosing potential learning difficulties related to lack of cognitive and metacognitive regulation.

Suggestion 6: Establish guidelines for implementing and evaluating instruction. Metacognitive theory has not been translated adequately into educational practice. This sentiment was captured well by Baker (1989), Garner (1987), and Baker and Cerro (this volume). Few teachers have a clear sense of how to improve metacognition and metamemory, or even ways to enhance the growth of specific subcomponents of metacognition such as conditional knowledge, although some educators are quite skilled in this regard. Educators desire guidelines for helping their students become more metacognitively aware. These guidelines should explicate separate subcomponents of metacognition (e.g., conditional knowledge, monitoring) as specifically as possible, and propose specific instructional interventions that improve these skills. Researchers also must propose guidelines for assessing the growth of metacognitive skills.

A number of researchers have investigated the kind of instructional interventions I am describing (Brown & Pressley, 1994; Delclos & Harrington, 1991; King, 1991; Paris & Jacobs, 1987; Palinscar & Brown, 1989; Zimmerman & Martinez-Pons, 1990). Many of these interventions have focused on teaching specific strategies (e.g., identifying main ideas) and monitoring their use. Others have proposed a broader research agenda (Garner, 1990; Pressley, Harris, & Marks, 1992; Van Meter, Yokoi, & Pressley, 1994). My own view is that instructional research in metacognition should proceed in several ways. One is the traditional theory-driven approach reflected in the work of Palinscar and Brown (1984). Another avenue is to utilize in-
class observations to construct grounded theories of metacognition (Strauss & Corbin, 1990; Pressley, this volume). Grounded theories seem especially important at this juncture given the lack of a unified theory of metacognition. A third approach is to utilize phenomenological methods (Creswell, 1994; Moustakas, 1994) that provide an in-depth descriptive account of what effective teachers do to improve metacognition.

Educators need to assess the effectiveness of their interventions as well. As I have suggested at several points in this chapter, guidelines and proven instruments for doing so are lacking. There is much that teacher-preparation programs could do to enhance future educators' knowledge of measurement and assessment. However, at present, many preservice teachers appear to be ill-prepared to meet assessment challenges. At a minimum, classes in reading, science, and mathematics instruction should include methods for providing a knowledge base, as well as metacognitive knowledge about regulating that knowledge base. Suggestions for assessing the effectiveness of both kinds of instruction (i.e., cognitive and metacognitive skills) should be embedded within this context.

Suggestion 7: Consider the relationship among metacognitive and affective variables. This suggestion follows from the work of Borkowski and colleagues (Borkowski & Muthukrishna, 1992; Borkowski, Millstead, & Hale, 1988; Borkowski et al., this volume) that describes the relationship among metacognitive and affective variables. Far less has been made of these important connections than they deserve (see Weinert, 1987). However, several studies suggest that metacognition may play a role in increasing personal interest and reducing anxiety that interferes with task performance (Tobias, 1995; Tobias & Everson, this volume). Similarly, metacognition may facilitate the understanding and regulation of emotions and possible selves in academic settings (Borkowski & Thorpe, 1994). Last, Pintrich and colleagues (Garcia & Pintrich, 1994; Pintrich & DeGroot, 1990; Pintrich et al., this volume) have elaborated on a number of important connections between metacognition and motivation.

Summary

This section has proposed a broad agenda for translating metacognitive theory into educational practice on a broad scale. This goal is ambitious. It will take the better part of a decade under the best of circumstances. The success of this research depends in large part on the ability of theorists, instructional design experts, and
specialists in educational measurement to carve out a common agenda. This agenda should accomplish the following: (a) provide operational definitions of specific metacognitive skills and their relationship to the student’s knowledge base, (b) construct and evaluate a variety of outcome measures that can be used to assess these skills, (c) cross-validate these outcome measures so that they can be used collectively to conduct multidimensional assessment of metacognitive skills, and (d) train practicing teachers to teach and evaluate metacognitive skills.

Conclusion

This chapter has attempted to provide an overview of salient measurement issues that are relevant to the study of metacognition. Although I believe the construct of metacognition is essential to understanding human cognition, the past two decades of research and practice have not achieved the lofty goal of presenting a comprehensive theory of metacognition that can be rendered into educational practice. It is time to take this goal seriously. The Buros Symposium on Issues in the Measurement of Metacognition was convened for this specific purpose. The chapters included in this volume have turned over many rocks, showing us the good, the bad, and the ugly.

Yet although a call to arms is warranted, it would be misleading to suggest that the field has reached an impasse. As evidenced by the chapters in this book, researchers agree on many essential points regarding metacognitive theory and practice. There is substantial agreement that the related constructs of metacognition and self-regulation have made an enormous contribution to cognitive psychology, literally changing the way that psychologists and educators view cognition and development. There also is widespread agreement that it is time to shore-up our knowledge of metacognition with sound measurement practice.

The separate and combined contributions of this volume point us in the right direction for substantial progress over the next decade. Metacognition has much to offer teachers and researchers. I believe the field of metacognition can deliver on its promise of helping students at all levels understand their thinking and learning.

REFERENCES


Author Index

A

Adrales, K.: 108, 110, 139
Afflerbach, P. P.: 126, 144
Ahmad, M.: 68, 95
Albertson, L. R.: 15, 36
Alexander, J. M.: 223, 229, 231, 251, 254
Alexander, P.: 45, 46, 66, 90, 103, 139
Algaze, B.: 221
Allen, B. A.: 114, 135
Almasi, J.: 294
Alvermann, D. E.: 117, 145
American Educational Research Association: 305, 315
American Psychological Association: 305, 315
Ames, C.: 4, 28, 29, 32, 33, 34, 35
Anand, P. G.: 199, 216, 220
Anderson, R. L.: 105, 136
Anderson, V.: 197, 218
Antes, J.: 105, 139
Anton, W. D.: 221
Archer, J.: 4, 28, 33, 35, 37
Armour-Thomas, E.: 114, 135, 226, 254
Artzt, A. F.: 226, 254
Ashford, S. J.: 202, 216
Assessment Systems Corporation: 244, 245, 254
August, D. L.: 104, 138
Ayres, R. R.: 21, 37

B

Baker, E.: 55, 93
Barker, K.: 267, 293
Barnes, J.: 22, 39
Baron, J.: 52, 91
Barr, R. G.: 204, 217
Baxter, G. P.: 43, 59, 60, 62, 88, 89, 91, 92, 95, 147, 298, 301
Belmont, J.: 11, 35
Belmont, J. M.: 103, 109, 130, 136
Bendixen, L.: 76, 95, 223, 234, 251, 259, 302, 320
Bendixen, L. D.: 14, 40, 212, 221
Benton, S. L.: 124, 136
Bereiter, C.: 52, 91, 99, 103, 136, 142
Bergman, J. L.: 294, 295
Betancourt, H.: 75, 91
Bird, M.: 103, 136
Bjorklund, D. F.: 2, 40, 227, 254
Bleiler, T.: 240, 260
Bohning, G.: 118, 143
Bollen, K. A.: 306, 316
Bondy, E.: 122, 126, 136
Borkowski, J.: 254, 298, 316
Borkowski, J. G.: 1, 3, 4, 5, 8, 9, 11, 12, 13, 14, 15, 18, 19, 26, 27, 29, 35, 36, 37, 38, 39, 40, 101, 102, 103, 106, 109, 124, 130, 136, 137, 140, 142, 149, 216, 223, 226, 258, 267, 294, 298, 299, 300, 305, 309, 310, 314, 319
Bosky, A. B.: 117, 145
Bower, G. H.: 264, 293
Boyle, R.: 52, 94
Bracken, B. A.: 199, 216
Branden, C. J.: 109, 142
Braford, J.: 45, 91
Bransford, J. D.: 15, 16, 31, 36
Brelend, H.: 152, 157, 209, 210, 216
Brennan, R. L.: 59, 92
Britten, T.: 267, 293
Britton, B. G.: 204, 217
Brophy, J.: 33, 36
Brown, A.: 45, 91, 100, 122, 137, 223, 224, 226, 255, 298, 303, 305, 310, 313, 316, 319
Brown, A. L.: 31, 32, 36, 40, 99, 100, 102, 103, 104, 129, 136, 137, 147, 185, 204, 216, 226, 254, 311, 316
Brown, J. S.: 32, 36
Brown, R.: 269, 294, 295, 296, 303, 310, 313, 316
Brown, W.: 78, 91
Bruce, B. C.: 123, 137
Brunner, J.: 122, 124, 141
Bruno, K.: 114, 135
Bulgarelli, C.: 153, 220
Burack, N. J.: 40
Burke, C.: 117, 139
Burke, K.: 120, 123, 137
Burland, S.: 108, 140, 149, 219
Burt, C. W.: 130, 140
Butler, D. L.: 202, 216
Byrd, D. M.: 130, 137
Byrnes, J. P.: 223, 257, 305, 316
Cairns, L.: 22, 39
Caldwell, J.: 126, 140
Cameron, R.: 108, 140, 149, 219
Campbell, D. T.: 58, 91, 308, 316
Campione, J.: 45, 91
Campione, J. C.: 31, 36, 185, 204, 216
Cardelle-Elawar, M.: 185, 217
Cariglia-Bull, T.: 130, 143
Carlo, M.: 149, 220
Carlo, M. S.: 177, 220
Carlson, J. S.: 207, 217
Carpenter, P. A.: 152, 157, 210, 218
Carr, M.: 3, 13, 19, 35, 36, 102, 124, 137, 223, 229, 251, 254
Cavanaugh, J. C.: 101, 109, 137, 140, 224, 255
Cerro, L.: 136, 301, 303, 307, 311, 312, 313
Cerro, L. C.: 99, 103, 137, 149, 236
Chan, L. K. S.: 1, 19, 20, 22, 23, 24, 25, 37, 38, 39, 41, 298
Charalambous, D.: 172
Cheung, P. C.: 29, 39
Chi, M. T.: 250, 252, 256, 300, 317
Chi, M. T. H.: 269, 293, 298, 316
Cholowski, K. M.: 25, 37
Chowning, F.: 118, 143
Cisero, C. A.: 149, 220
Clayton-Jones, L.: 19, 23, 37
Clifford, M. M.: 18, 37
Cobb, P.: 27, 29, 37
Cognition and Technology Group at Vanderbilt: 16, 31, 37
College Board: 153, 158, 217
College Entrance Examination Board: 176, 177, 217
Collier, G. V.: 127, 139
Collins, A.: 32, 36
Commeyras, M.: 123, 137
Connell, M. L.: 185, 216
Cook, T. D.: 308, 316
Cooley, E. J.: 21, 37
Cordon, L. A: 263, 290, 293
Corkill, A. J.: 304, 317
Corno, L.: 52, 91
Craik, F. I. M.: 215, 217
Creelman, C. D.: 157, 219
Cronbach, L. J.: 53, 91, 158, 217
Cross, D.: 65, 92, 93
Cummings, L. I.: 202, 216
Cunicelli, E. A.: 40
Cutler, B. L.: 234, 255
Cutler, S. K.: 204, 222

D

D'Agostino, A.: 187
Davis, M. A.: 192, 219
Day, J. D.: 9, 15, 53, 37, 101, 137, 263, 290, 293
de la Torre, R.: 240, 260
debus, R. I.: 22, 39
Deci, E. L.: 4, 34, 37, 196, 217
Deegan, D. H.: 268, 293
DeGroot, E.: 51, 58, 78, 80, 83, 84, 94, 301, 310, 314, 319
DeGroot, E. V.: 101, 102, 103, 112, 142
Delaney, H. D.: 198, 217
Delclos, V. R.: 231, 232, 255, 303, 313, 317
Dennison, R.: 62, 63, 64, 95
DeVito, P. J.: 126, 143
Diener, E.: 234, 259
Diemeyer, D.: 9, 37, 101, 137
Dings, J.: 240, 260
Dixon, R. A: 108, 130, 139, 226, 255
Douglas, V. T.: 204, 217
Duffy, S. A.: 267, 294
Dufresne, R.: 177, 220
Duguid, P.: 32, 36
Dunbar, S.: 55, 93
Dunkle, M.: 76, 95
Dunkle, M. E.: 14, 40, 212, 221, 223, 234, 251, 259, 302, 320
Dunning, D.: 234, 255
Dunning, T.: 123, 138
Dweck, C. S.: 28, 34, 38

E

Earthman, E. A.: 268, 293
Ee, J.: 25, 38
Einstein, G. O.: 235, 257
Ekstrom, R. B.: 174, 217
El-Dinary, P. B.: 269, 294, 295, 296
Elder, A. D.: 89, 91
Elig, T. W.: 21, 38
Elliot, E. S.: 34, 38
Elliott-Faust, D. J.: 2, 40
Ellis, E. S.: 116, 124, 125, 138
Erickson, L. G.: 106, 138
Ericsson, K.: 83, 92
Estrada, T. M.: 101, 137
Evans, P.: 269, 296
Everson, H.: 67, 68, 72, 75, 96, 147, 152, 157, 222, 300, 308, 312, 314
Everson, H. T.: 150, 212, 217, 218
Eysenck, M.: 201, 218

F

Faculties and Administration of Benchmark School: a: 295
Fairbanks, M. M.: 118, 142
Farr, M. J.: 269, 293
Farrell, P.: 45, 91
Fenema, E.: 191, 218
Ferrara, R.: 45, 91
Fischer, P. M.: 103, 138
Fischoff, B.: 234, 255
Fiske, D. W.: 44, 45, 46, 66, 92, 147, 218
Flower, L.: 267, 293
Fodor, J.: 303, 317
Foley, J.: 123, 138
Foley, M. J.: 227, 257
Foorman, B. R.: 230, 234, 256
Forlizzi, L.: 136
Forrest-Pressley, D. L.: 2, 40, 99, 138
Fox, P. W.: 229, 257
Frederiksen, C. H.: 268, 293
Frieze, I. H.: 21, 38
Funder, D. C.: 233, 255
Furando, J.: 136

G

Gao, X.: 59, 62, 88, 92, 95
Garcia, M. W.: 127, 138
Gardner, H.: 303, 317
Garofalo, J.: 185, 219
Gaskins, I.: 294
Gaskins, I. W.: 4, 40, 295
Gavelek, J. R.: 27, 40
Geary, D. C.: 108, 110, 139
Gelman, S. A.: 303, 318
Gelzheiser, L.: 12, 38
Gerler, D.: 67, 93, 214, 219
Gerrity, H.: 173
Ghatala, E.: 48, 49, 50, 56, 68, 73, 74, 75, 92, 94, 95
Ghatala, E. S.: 192, 220, 223, 224, 227, 228, 229, 230, 231, 234, 251, 256, 258, 263, 292, 295, 310, 319
Gholson, B.: 130, 137
Glaser, R.: 89, 91, 124, 139, 250, 252, 256, 269, 293, 300, 317
Glover, J. A.: 103, 142
Goertz, M. E.: 174, 217
Golan, R. M.: 34, 38
Gonzalez, H. P.: 221
Goodman, Y.: 117, 139
Gourgey, A.: 68, 96, 150, 152, 212, 217, 218, 222
Grabe, M.: 105, 139
Graesser, A. C.: 264, 293
Graham, S.: 34, 38, 66, 75, 89, 90, 92
Graves, B.: 268, 293
Gray, M. J.: 117, 119, 139
Green, D. M.: 157, 218
Green, F. L.: 104, 138
Gridley, B.: 63, 93
Gridley, B. E.: 107, 111, 140
Griffin, D. W.: 234, 255
Grolnick, W. S.: 35, 38
Groteluschen, A.: 101, 137
Gruson, L.: 108, 140, 149, 219
Guba, E. G.: 266, 293, 294
Guthke, J.: 207, 218
Guthrie, J.: 267, 293
Guthrie, J. T.: 298, 319
H
Haas, C.: 267, 293
Hahn, H.: 105, 139
Hale, C.: 14, 36, 314, 316
Hansen, J.: 126, 139
Hare, V.: 45, 90
Hare, V. C.: 267, 293
Harnishfeger, K. K.: 227, 254
Harris, K.: 82, 86, 94
Harris, K. R.: 31, 40, 227, 258, 298, 303, 313, 319, 320
Harris, L. J.: 176, 218
Hart, J. T.: 214, 218
Hartman, H.: 68, 96, 150, 152, 212, 217, 218, 222
Hassebring, T. S.: 16, 36
Hastroudi, F.: 224, 256
Haynes, N. M.: 114, 135
Heath Reading Strategies Assessment: 126, 139
Hembree, R.: 190, 195, 218
Herndon, J. N.: 199, 218
Hertzog, C.: 108, 130, 139, 226, 255
Hidi, S.: 196, 197, 218, 220
Hiebert, E. H.: 126, 144
Hirschfeld, L. A.: 303, 318
Holen, M. C.: 175, 218
Holst, P. M.: 236, 258
Holtzman, W.: 78, 91
Hotelling, H.: 166, 218
Howard-Rose, D.: 63, 83, 87, 88, 92, 300, 318
Howespian, B. A.: 9, 37
Hultsch, D. F.: 108, 130, 139, 226, 255
Human, P.: 27, 39
Hunter-Blanks, P.: 48, 68, 70, 73, 75, 92
Hutchings, C. H.: 192, 219

I
Indrisano, R.: 117, 125, 141
Irwin, D. E.: 249, 257

J
Jacobs, J.: 61, 62, 63, 64, 65, 92, 93
Jacobs, J. E.: 101, 103, 107, 108, 110, 111, 117, 124, 130, 139, 141, 150, 218, 226, 256, 303, 310, 313, 318
Jenkins, J. J.: 262, 294
Jenkins, L.: 152, 157, 209, 210, 216
Jensen, A. R.: 251, 256, 305, 318
Johnson, M. K.: 224, 256
Johnson, P. L.: 240, 260
Johnston, J. C.: 15, 36
Johnston, P.: 102, 103, 135, 144
Jones, R. J.: 152, 157, 209, 210, 216
Joreskog, K. G.: 24, 38, 165, 167, 222
Just, M. A.: 152, 157, 210, 218

K
Kahneman, D.: 234, 259
Kajer, W. K.: 227, 257
Kaniel, S.: 125, 139
Kapinus, B. A.: 127, 139
Karmiloff-Smith, A.: 303, 318
Keefer, K. E.: 175, 218
Kennedy, M. L.: 118, 142
Keren, G.: 229, 233, 256, 312, 318
Khan, M.: 227, 260
Kiewra, K. A.: 124, 136
King, A.: 303, 311, 313, 318
Kintsch, W.: 264, 294
Kinzer, C. K.: 16, 36
Kirby, J. R.: 108, 110, 133, 140
Kirby, K.: 118, 141
Kistner, J. A.: 18, 21, 38, 39
Klosterman, I. H.: 108, 110, 139
Kluwe, R. H.: 224, 226, 256, 260
Koriatt, A.: 49, 92, 214, 219, 224, 228, 234, 251, 252, 256, 310, 318
Korkel, J.: 109, 143
Koshida, D. T.: 304, 317

Pataschnick, M.: 28, 29, 39

Pearson, P. D.: 124, 127, 131, 144

Peck, V.: 229, 255

Perlmutter, M.: 101, 137, 224, 255

Perlwitz, M.: 37

Petty, C. R.: 214, 220

Phifer, S. J.: 103, 142


Pintrich, P. R.: 3, 14, 43, 45, 46, 51, 52, 58, 78, 79, 80, 83, 84, 87, 92, 94, 96, 101, 102, 103, 112, 142, 147, 149, 150, 206, 220


Place, N. A.: 126, 144

Plake, B. S.: 240, 252, 258, 260

Polin, R.: 138

Pollack, J. M.: 174, 217

Potenza, M.: 229, 259

Powell, H.: 40

Powell, J. L.: 126, 142

Presidential Task Force on Psychology in Education: 128, 142

Pressley, G. M.: 263, 294


Pridemore, D. R.: 230, 259

Puncochar, J.: 229, 257

R

Rakliffe, G.: 138

Ramist, L.: 165, 167, 222

Randall, A.: 118, 142


Reder, L. M.: 214, 220

Reichenberg, R.: 125, 139

Reid, M. K.: 11, 12, 40, 109, 140

Relich, J.: 22, 39

Rellinger, E.: 102, 137

Rellinger, L.: 3, 35

Remminger, K. A.: 196, 197, 220

Rhodes, L. K.: 112, 117, 120, 142

Rice, D. N.: 118, 140

Rinehart, S. D.: 106, 138

Ritter, F. E.: 214, 220

Rock, D. A.: 174, 217


Rocklin, T. R.: 236, 243, 244, 245, 250, 252, 258, 259

Rodwell, K.: 37

Roedel, T.: 76, 95


Roehler, L. R.: 138

Reeser, R.: 84, 94

Rogers, H. J.: 237, 256

Roos, L.: 298

Roos, L. L.: 223, 240, 241, 244, 245, 252, 258, 259, 260

Rosado, D.: 185

Ross, G. R.: 199, 221

Ross, L.: 234, 255

Ross, S. M.: 216, 220

Royer, J. M.: 149, 153, 177, 220

Ruiz-Primo, M. A.: 60, 95

Rushton, J. P.: 109, 129, 142

Ryan, M. P.: 214, 220

Ryan, R. M.: 4, 34, 35, 37, 38, 196, 217


S

Saenz, D.: 101, 137
Saenz, D. S.: 9, 37
Sanocki, T.: 224, 227, 256
Sarason, I. G.: 201, 220
Scardamalia, M.: 52, 91, 99, 142
Schallert, D.: 45, 90
Schaub-Matt, M.: 40
Schiefele, D.: 196, 197, 221
Schmitt, M. C.: 112, 143
Schneider, W.: 2, 3, 14, 40, 52, 95, 101, 109, 142, 143, 224, 226, 258, 259, 267, 294, 299, 319
Schoenfeld, A. H.: 31, 32, 40
Schoenfeld, A. H.: 158, 191, 193, 194, 221, 240, 259
SPSS: 168, 221
Squires, J. R.: 277, 295
Squires, P.: 136
Stahl, S. A.: 106, 138
Stein, B. S.: 15, 36
Stein, S.: 269, 296
Steinberg, I.: 118, 143
Sternberg, R. J.: 101, 113, 114, 124, 143, 221, 305, 320
Stevens, D. D.: 27, 40
Stipek, D.: 21, 22, 41
Stock, W. A.: 230, 259
Suh, S.: 267, 295
Swaminathan, H.: 237, 256
Swanson, H. L.: 114, 143, 204, 221, 226, 231, 251, 259, 303, 320
Swanson, L.: 65, 96
Swets, J. A.: 157, 218
Symeons, S.: 130, 143

T

Tanner, T.: 173, 221
Taylor, C. J.: 221
Taylor, T. R.: 124, 125, 128, 132, 143
Terwogt, M. M.: 106, 139
Thomas, D. L.: 234, 259
Thompson, R. C.: 227, 257
Thorpe, P.: 26, 36
Thorpe, P. K.: 298, 305, 314, 316
Thorson, I.: 105, 139
Tindal, G. A.: 312, 320
Tobias, S.: 45, 67, 68, 69, 71, 72, 75, 78, 96, 147, 150, 152, 153, 157, 190, 195, 196, 197, 198, 201, 208, 212, 215, 217, 218, 221, 222, 229, 260, 300, 308, 312, 314, 320
Torgesen, J. K.: 108, 115, 128, 131, 133, 143
Touchstone Applied Sciences Associates: 204, 222
Trabasso, T.: 267, 295
Tracy, A.: 138
Trahan, M. F.: 204, 221
Trathen, W.: 267, 295
Traweek, D.: 103, 108, 124, 125, 144
Trigatti, B.: 37
Tuckman, B.: 202, 222
Turner, J. C.: 99, 141
Turner, L. A.: 18, 36
Tversky, A.: 229, 233, 234, 257, 259

V

Valencia, S.: 124, 126, 127, 131, 144
Valencia, S. W.: 126, 128, 144
van den Broek, P.: 264, 295
Van der Westhuizen, G.: 116, 141
Van Haneghan, J. P.: 99, 104, 144, 193, 222
Van Meter, P.: 313, 320
Vavrus, L.: 138
Verville, K.: 127, 138
Vispoel, W. P.: 240, 241, 260
Visser, P.: 106, 139

W

Wade, S. E.: 267, 295
Waller, T. G.: 99, 138
Wang, T.: 240, 260
Ward, L.: 103, 108, 124, 125, 144
Wasik, B. A.: 99, 116, 141
Watson, D.: 117, 139
Weaver, C. A., III: 224, 226, 227, 234, 259, 260
Weiner, B.: 17, 21, 41
Weinert, F. E.: 109, 143, 224, 260, 314, 320
Wellman, H. M.: 1, 38
Wellman, J. M.: 2, 41
Wentzlaff, J. R.: 214, 220
Werner, H.: 81, 96
Werts, C.: 165, 167, 222
Westberry, L. G.: 221
Westby, C. E.: 204, 222
Weyhing, R. S.: 13, 18, 19, 36
Wheatley, G.: 37
Wiedl, K. H.: 207, 208, 217
Wigfield, A.: 21, 41
Willert, M. G.: 227, 257
Williams, S. M.: 16, 36
Willingham, W. W.: 165, 167, 222
Willoughby, T.: 227, 260
Wilson, J.: 204
Wineberg, S. S.: 268, 296
Winne, P.: 63, 83, 87, 88, 92, 227, 260, 300, 318
Winne, P. H.: 202, 216
Winograd, P.: 34, 40, 44, 93, 101, 102, 142, 144
Winser, B.: 103, 104, 118, 121, 144
Wise, S. L.: 223, 240, 241, 243, 244, 245, 250, 252, 253, 258, 259, 260, 298
Wittrock, M. C.: 108, 133, 145
Wixson, K.: 46, 93
Wixson, K. K.: 117, 127, 145
Wolfe, R. N.: 234, 255
Woloshyn, V.: 50, 94, 230, 258, 263, 295
Wolters, C. A.: 43, 80, 96, 147, 298, 301
Wood, E.: 227, 260
Wood, T.: 37
Wright, C. D.: 199, 222
Wright, J. P.: 199, 222
Wyatt, D.: 269, 276, 277, 290, 293, 296
Y
Yackel, E.: 37
Yan, W.: 251, 260
Yates, J. F.: 229, 260
Yochum, M. N.: 117, 145
Yochum, N.: 117, 118, 120, 129, 145
Yoes, M. E.: 244, 245, 259
Yokoi, L.: 313, 320
York, M.: 295
Youlden, A. M.: 25, 41

Z
Zabrucky, K.: 105, 106, 107, 117, 118, 119, 120, 128, 141, 145
Zimlin, L.: 106, 136
Zimmerman, B. J.: 45, 50, 51, 52, 58, 71, 78, 79, 81, 82, 84, 85, 87, 95, 96, 97, 298, 313, 320, 321
Zimmerman, S.: 58, 96, 301, 320
Zimmerman, S. A.: 103, 113, 144
### Subject Index

| A | Cognitive capacity: 191 |
| allocation of resources: 52 | cognitive skills: 226, 301, 314 |
| 243, 252, 255, 287, 301, 314 | communities of learners: 31, 33 |
| assessment | comprehension |
| authentic: 120-122, 124, | standards of evaluation: |
| 126, 290 | 106-107, 119 |
| by practitioners: 86, 129, | comprehension monitoring: 48, 51, |
| 134 | 102, 105-107, 118-120, 224, 263, |
| by teachers: 59, 66, 76, 85, 87, | 288, 303, 310 |
| 93, 116-117, 119-123, 129 | comprehension strategies: 263-264, |
| dynamic: 124-125, 135 | 275 |
| formal: 76, 112, 120 | computer-based testing: 223-254, |
| informal: 76-77, 112, 116- | 308 |
| 118, 120 | computerized adaptive testing |
| metacognitive knowledge: | (CAT): 239-241, 248, 252-253 |
| 148, 301-302, | concurrent validity: see validity |
| 306, 312 | conditional knowledge: 45, 63, 226, |
| performance: 105, 124, 135, | 313 |
| 290, 297, 306, 308-309, | construct validity: See validity |
| 311 | constructivism: 29 |
| portfolio: 111, 125-126, 135 | control: 224-227, 231-232, 235-236, |
| attention deficit hyperactivity disorders: 204-206 | 241, 244, 249-254, 298, 305, 308, |
| attributional beliefs: 3, 7-8, 10-11, | 310 |
| 17-21, 25-26, 102 | converging evidence: 129, 135 |
D
declarative knowledge: 45, 63, 203, 206, 212, 223, 226
diagnostic testing: 115, 119, 123
domain generality: 59, 76, 78, 84, 109, 115, 130, 135, 212–213, 228, 235, 250, 252, 302, 303
domain knowledge: 229, 231, 252, 269
beliefs: 7, 10
knowledge: 7, 9–10
dynamic assessment: 207, 209

E
ease of learning judgments: 48, 67, 71, 73, 76–77, 229
ecological validity: 107, 114, 121, 131, 135, 311
error detection: 50, 68, 70, 73–75, 77, 103–107, 116, 118–122, 134, 236, 312
estimates of academic achievement: 180, 185–186
estimates of knowledge: 159, 207, 209–214
executive processing: 2–4, 7–8, 13–16, 20, 191, 204, 298

F
feedback: 67, 201–203, 206–207, 210, 212, 228, 230, 235, 252, 313
feeling of knowing: 214–215, 229

G
generalizability: 59, 62, 65–67
generalization: 2, 12–13, 15, 66, 83
“good information” processing: 2–5, 9

grade point average: 165–173, 175–176, 193–194

I
intrinsic motivation: 4, 10, 84, 196
introspective reports: 149
item response theory (IRT): 225, 236–240, 244, 245

J

K
knowledge-base specific: 79
knowledge construction: 3, 30–31
knowledge of cognition: 226, 301

L
learning and study strategies: 109, 112–113, 115, 128, 134, 231, 301
learning contexts: 26–29, 32

M
Subject Index


mental knowledge: 1, 3
mental models: 15

metacognition
development of: 1–9, 27
relationship to performance: 26
reliability: 2, 16
validity: 2, 11, 21

awareness: 111, 113, 115, 125–126, 133, 226
development: 100, 251, 301
judgments: 44, 47–48, 67–68, 73, 75
measurement: 223–254, 296, 307, 312
regulation: 261, 302, 304, 313
skills: 226, 297, 303, 313–315
metacognitive model: 8–11, 17
metacomponents: 113–114, 124
metamemory: 1–2, 12, 49, 73, 75, 77, 109, 113, 124, 132, 134, 214, 313
motivation: 2–4, 8, 17, 22, 28, 45, 50, 52, 83–84, 90, 112, 131, 196, 215, 242, 287, 289, 299, 301, 305, 314
motivational beliefs: 52

O
observation: 148, 270
on-line process measures: 105–106, 128

P
performance estimation: 176, 183–184, 207
performance judgments: 229
performance monitoring: 298, 310
possible selves: 4–5, 9, 314
predictions of performance: 178, 180, 184, 189, 210
procedural knowledge: 48, 63, 151, 185, 187–188, 190, 206, 223, 226
process measures: 106, 113, 116, 120, 125, 128, 134
protocol analysis: 76
Subject Index

**Q**

questionnaires: 60, 66, 78, 84–86, 110, 150

**R**


reading behaviors: 268, 271, 275


reading protocols: 271

reading strategies: 270–271, 275

reciprocal teaching: 32–33

regulation of cognition: 45, 50–51, 62, 78, 83, 100–101, 113, 224, 226, 302

regulatory control: 298, 310

reinforcement: 201–202, 203


**S**

school learning: 151, 153, 159–160, 166–167, 175, 183, 185, 209, 215


self-assessment: 128

self-regulated learning: 43–90


signal detection theory: 157

situational interest: 160–161, 197–199

strategy effectiveness: 121

strategy instruction: 227, 313

strategy knowledge: 2–8, 10–11, 56, 160

strategy revision: 13, 16


strategy variables: 46

**T**

task analyses: 7, 16

task monitoring: 298, 310

task specificity: 59, 64, 78, 109

task variables: 46

task vs. ego orientations: 28–29

test anxiety: 84, 240


topic interest: 197–198

training: 76, 148, 208, 230–231, 299, 303–304, 313

**U**


**V**


volitional control: 52