2000

2. Assessing Metacognition and Self-Regulated Learning

Paul R. Pintrich
University of Michigan, Ann Arbor

Christopher A. Wolters
University of Houston, cwolters@uh.edu

Gail P. Baxter
Educational Testing Service

Follow this and additional works at: http://digitalcommons.unl.edu/burosmetacognition

Part of the Cognition and Perception Commons, and the Cognitive Psychology Commons

http://digitalcommons.unl.edu/burosmetacognition/3

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.
 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-
cal 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 comprehen-
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 weaknesses 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 \textit{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 \textit{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 \textit{when} and \textit{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.

<table>
<thead>
<tr>
<th>Evidential Basis</th>
<th>Test Interpretation</th>
<th>Test Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell 1</td>
<td>Construct Validity</td>
<td>Cell 2</td>
</tr>
<tr>
<td>1) Content</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2) Substantive</td>
<td>Relevance/Utility</td>
<td></td>
</tr>
<tr>
<td>3) Structural</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4) External</td>
<td>Value Implications</td>
<td></td>
</tr>
<tr>
<td>5) Generality of Meaning</td>
<td>Social Implications</td>
<td></td>
</tr>
</tbody>
</table>

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 construct­irrelevant 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.
2. ASSESSMENT OF METACOGNITION

CONSTRUCT VALIDITY OF MEASURES OF METACOGNITION
AND SELF-REGULATED LEARNING

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 & Afflerbach, 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 unrecalled 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, Ghatatala, Pressley, & Levin, 1988; Pressley, Snyder, Levin, Murray, & Ghatatala, 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 perfor-
2. ASSESSMENT OF METACOGNITION

Performance 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
Tobias, this volume) that can be more easily used in classrooms by researchers and teachers.

**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
2. ASSESSMENT OF METACOGNITION

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

In J. Kuhl & J. Beckman (Eds.), *Action control: From cognition to behavior* (pp. 101-128). Berlin: Springer-Verlag.


