7. Assessing Metacognition: Implications Of The Buros Symposium

Gregory Schraw
University of Nebraska-Lincoln

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

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

https://digitalcommons.unl.edu/burosmetacognition/8

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

Six Emergent Themes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**SUMMARY**

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

**Thoughts from the Measurement Community**

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

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

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

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

Both Flavell (1987) and Brown (1987) proposed descriptive theories of metacognition that, although overlapping, remain somewhat independent of one another. Surprisingly, very few researchers have attempted to translate these theories into operational models that enable researchers to investigate systematically the relationships among model components. Two important exceptions are the models proposed by Nelson and Narens (1990, 1994) and Borkowski and colleagues (Borkowski & Muthukrishna, 1992; Borkowski & Thorpe, 1994; Borkowski et al., this volume). These models enable researchers to make explicit predictions about the relationships among control and monitoring processes, one’s extant knowledge base, and motivational factors that affect self-regulation.

Process models clearly play an important role in theory testing. Models specify particular relationships that can be tested explicitly. Of equal importance, models focus our attention on the role that instrumentation and measurement play in theory testing. The advent and continued growth of structural equation modeling, for example,
SCHRAV has ushered in a new era in the measurement of salient constructs in cognitive psychology. Such models require researchers to test not only the structural aspects of the model (i.e., relationships among model components predicted by theory), but measurement aspects as well (Bollen, 1989).

Caution 3: Construct and evaluate instruments that assess specific components of the model. Three questions seem especially germane to this caution. These include: (a) What is the construct of interest? (b) Do appropriate measures exist to measure it? and (c) Does the measurement process change the construct? Each of these questions is considered briefly.

What is the construct? Broad constructs are difficult to measure, making it unlikely that there will ever be a single measure of metacognition (see Pintrich et al., this volume; Schraw, 1995). An alternative approach would be to partition the metacognition construct into smaller components, then hypothesize about the relationships among subcomponents, although others may disagree with this strategy (Linn, 1991). Constructing a detailed structural model that specifies individual components and their interrelationships is an essential precursor to the validation process.

Do appropriate measures exist? Once potentially measurable subcomponents have been defined, researchers must select behavioral and self-report indices that measure these components. Indices of metacognitive activity should be evaluated in an ongoing manner using a variety of approaches (see Crocker & Algina, 1986, chapter 4, for a summary of this process). Potential measures need to be evaluated with respect to reliability, validity, and utility, or their qualitative counterparts such as credibility and authenticity (Creswell, 1994; Merriam, 1988; Miles & Huberman, 1984).

Two aspects of reliability are essential. One concerns the internal consistency of an instrument. Measures such as coefficient alpha provide easy-to-compute indices of internal consistency and are available for a variety of measures of metacognitive knowledge and performance (Pintrich et al., 1993; Schraw & Dennison, 1994; Weinstein et al., 1988). However, when used in isolation, coefficient alpha is insufficient. A second aspect of reliability is consistency over time. Currently, there is little available information regarding test-retest reliability on commonly used measures of self-regulation and metacognition.

Multiple aspects of validity are crucial to the effective use and interpretation of scores from an instrument or experimental results. In their chapter, Pintrich et al. (this volume) addressed the question of
validity using Messick's (1989) five-component framework. In this view, all questions of validity are essentially questions related to construct validity (i.e., the degree to which inferences about a score accurately represent an observable or unobservable phenomenon of interest). Given that Messick views each of these five components as interrelated, threats to any component necessarily affect all others.

A similar argument can be made for the credibility of qualitative methods. Whether researchers use terms such as utility or confirmability is beside the point; rather, they must demonstrate that they have explained a hypothetical construct (or phenomenon) in an accurate, replicable way. Researchers with a predilection for either qualitative or quantitative methods should conduct carefully planned validation studies to assure their methods accurately describe the phenomenon of interest. In particular, there is a tremendous need for studies evaluating the convergent and divergent validity of scores from multiple measures, preferably using multiple methodologies, to explain metacognitive phenomena. Point 3 made by Pintrich et al. (this volume), in which they recommended the use of a multitrait, multimethod approach to construct validation, emphasized this concern directly.

**Does the measurement process change the construct?** Researchers must ask themselves to what extent their measures of metacognition affect the deployment of metacognitive knowledge and regulatory skills. Although there are a variety of potential measurement confounds, several that I consider to be especially serious are discussed below (see Baker & Cerro, this volume). The first is that self-report inventories (e.g., Schraw & Dennison, 1994) may elicit socially desirable responses. One way to safeguard against instrument bias is to conduct convergent and divergent validity studies. Unfortunately, few studies of this kind have been reported.

A second potential problem is that think-aloud studies may affect the measurement of metacognitive processes by competing for limited resources that are necessary for task performance. One option is to use retrospective self-reports; however, verbal report theorists generally view retrospective reports as less reliable than concurrent self-reports. Competition for limited resources also affects many quantitative studies, but especially those using on-line confidence judgments of performance. One alternative is to use unobtrusive measures such as computerized testing procedures (Schraw et al., this volume).

A third problem is that structured interviews may provide information to individuals that they would not report on their own, thereby masking their true metacognitive knowledge. One option is
to ask informants to respond in writing, although this leads to other problems such as competition for limited resources.

A fourth problem, and the most serious in my opinion, is that individuals may differ in terms of their ability to explain or estimate their metacognitive knowledge either verbally or in writing. For example, experts may have a richer vocabulary for describing their mental processes, even though novices engage in similar processes. In essence, this problem is due to aptitude by treatment interactions, and must be considered carefully when researchers evaluate potential threats to the validity of their findings.

Caution 4: Use diverse assessment methods. Pintrich et al. (this volume) captured the main theme of the Buros Symposium when they concluded there is no one-size-fits-all measure of metacognition. Indeed, if we accept the premise that there are two or more distinct metacognitive processes (e.g., control and monitoring), it seems reasonable to conclude that there must be different measures of these processes. Equally reasonable is the assumption that each identifiable metacognitive process can be measured using different instruments and methodologies (see Borkowski et al., this volume).

Perhaps the best advice one could give researchers interested in metacognition is to adopt a multitrait, multimethod (MTMM) model of assessment along the lines first proposed by Campbell and Fiske (1959), and elaborated upon by Cook and Campbell (1979). The MTMM approach emphasizes the collection of multiple measures of a phenomenon using multiple, preferably diverse, methodologies. In the context of validating metacognitive theory, this means using objective self-reports (Pintrich et al., this volume), subjective self-reports such as concurrent verbal reports (Pressley, this volume), subjective assessments of one's thinking or performance such as calibration judgments (Tobias & Everson, this volume), and unobtrusive measures such as item selection times using computer-based testing (Schraw et al., this volume). Other measurement approaches are needed as well, such as neurophysiological correlates of metacognition (Metcalfe, 1994b).

Summary

The four cautions outlined above summarize basic measurement concerns for any research agenda. They have special importance for the domain of metacognitive research given the paucity of systematic validation studies. Foremost, the field needs a systematic and comprehensive assessment agenda, lest it lose its credibility among psy-
chologists and educators in general, and measurement experts in particular. A major part of this agenda should include the development of a testable model of metacognition. Work in this direction was described by Borkowski et al. (this volume). Second, working models must be tested and refined in an ongoing basis. Earlier stages of this research will undoubtedly focus on specific components of metacognition, whereas the later stages may test multicomponent models using procedures such as structural equation modeling; that is, the grain size of these studies would be expected to change over time. Third, a variety of measures should be used to assess each separate component, as well as the entire model. Last, adopting a multitrait, multimethod approach to model testing may enhance the effectiveness of this research.

An Agenda for Future Research

There are many ways that the assessment of metacognition can be improved in the future, but it is not an easy task, and it cannot be done without the help of colleagues in different fields of study. This section makes seven suggestions for future research and practice based on the cautions described above. I have rank ordered these suggestions in a way that might surprise some readers. Regardless of their ordering, all are essential to the advancement of the field, and should be taken seriously.

Suggestion 1: Researchers interested in metacognition must collaborate with measurement and instructional design experts. Most of the authors included in this volume would acknowledge their lack of technical measurement expertise. Most of them, and most researchers working in the area of metacognition, are not keenly interested in measurement issues per se. Many, including myself, do not have a strong interest in the design of instructional interventions. Yet it is clear to me that failing to involve measurement and instructional design experts may be disastrous to the field. I believe we are at an impasse that cannot be overcome without the skills and knowledge of experts who do not share our vested interest in the construct of metacognition.

On a brighter note, let me focus on some of the potential advantages of cross-disciplinary collaborations. One is that experts at solving measurement problems may add a tremendous amount of richness to existing theory. A second is that measurement and instructional design experts may provide innovative ways to assess metacognition, both with respect to objective paper-and-pencil measures, as well as performance-based assessments. A third is that
instructional design experts have much to offer by way of translating metacognitive theory into instructional practice. The same may be said regarding how to evaluate formative and summative instructional outcomes.

Suggestion 2: *Agree on a unified theoretical framework.* It bears repeating that the field is perceived by some outsiders as too theoretically disparate. It is unlikely that substantially more progress will be made until researchers agree on what it is they are looking for. My personal preferences are for a three-component model of metacognition that emphasizes the role of regulatory control, performance monitoring, and task-monitoring processes. Such a model is already consistent with much of the theorizing being done in metacognition and self-regulation (Baker, 1989; Borkowski et al., this volume; Garner, 1987; Garner & Alexander, 1989; Jacobs & Paris, 1987; Nelson & Narens, 1994; Pintrich & DeGroot, 1990; Pressley & Afflerbach, 1995; Schraw & Moshman, 1995). Even earlier theories such as those proposed by Brown (1987) and Flavell (1987) that postulate two main components, are highly consistent with the three-component view.

It is my view that the most important issue separating existing accounts of metacognition is terminological differences, not assumptions about basic metacognitive processes. In a way, it is as if the field has agreed on what kind of pizza it wants to eat, but still can’t decide how to slice the pieces! I believe we are much closer to a comprehensive theory of metacognition than most casual observers would give us credit for. Perhaps our greatest challenge is to recognize the perception of disarray from outside the field and to resolve it!

Another theory-related suggestion is to focus on measuring the practical and statistical relationships among the three main components described above. Recent work by Nelson and Narens (1990, 1994) and Koriat (1993, 1994) have raised extremely important questions about the relationship between monitoring and control. The importance of this relationship, as well as ways to test it, were echoed by Schraw et al. (this volume). Borkowski et al. (this volume) also emphasized the relationship between performance and task monitoring.

Suggestion 3: *Identify suitable outcome measures that can be used as criteria to evaluate metacognitive behaviors.* Most of the work done in the field of metacognition has been devoted to generating theory. A sizable amount of work also has been done on testing a narrow band of metacognitive processes, but especially comprehension monitoring (See Baker, 1989; Pressley & Ghatala, 1990; and Schraw & Moshman, 1995, for reviews). Some work, but not a great deal, has been done on instructional improvement (Brown & Pressley, 1994; Jacobs & Paris,
Yet virtually no attention has been paid to identifying suitable outcome measures of metacognition, other than as they apply to testing components of metacognitive theories, or evaluating the efficacy of instructional interventions.

Of particular concern are measures with a high degree of ecological validity, including performance-based measures that capture mid-range (e.g., 3 months) and long-term (e.g., 6 months or more) development of metacognition. There are not any suitable methods in this regard at the present time, and unfortunately, few researchers seem to have considered this problem (see Baker & Cerro, this volume, and Garner, 1987, for further discussion).

Outcome measures should be identified that meet a variety of needs. One especially important need is to parallel measures of metacognition that are suitable for children and adults. Without such measures, comparisons among age groups at a single time are compromised, as are across-time comparisons within the same group. A second need is to identify measures that are suitable for field settings. Many studies rely on either checklists or overt behavioral measures. Lacking is a method for identifying multiple levels of metacognitive activity via self-report. A third need is to construct specific versus broad measures of metacognition. Most studies focus on monitoring judgments made prior to or after completing a test item. Although monitoring reflects an important component of metacognition, it fails to provide information about strategies for planning, debugging, or evaluating.

Another important question to ask about these measures is the degree to which they intrude on metacognitive processes. Asking individuals to think-aloud as they perform a task, for example, may interfere with that task by consuming limited resources, or it may facilitate metacognitive behaviors by calling explicit attention to task demands.

Suggestion 4: Empirically investigate the relationship among different outcome measures. One consequence of the small amount of work that has gone into identifying alternative outcome measures is that very little is known about the relationship among these measures. For example, there has not been a systematic comparison of the relationship between ratings of monitoring accuracy and self-reported strategies, even though these measures are used frequently in the literature. As measurement experts would remind us, the more measures we use to assess an outcome, and the more we know about the relationship among these measures, the better able we are to make reliable and
valid inferences about the construct under study. Put simply, it is crucial to converge on a construct using multiple, triangulated methods.

Suggestion 5: Establish methodological and data-analytic guidelines for measuring metacognition. Any research tradition quickly establishes its own informal or formal guidelines for administering and interpreting a task. One example is the use of monitoring accuracy judgments that pervade the metacognition literature. Nevertheless, with the exception of monitoring accuracy judgments (see Tobias & Everson, this volume), error-detection (see Baker & Cerro, this volume), and self-reported strategy use (see Pressley, this volume), there are few guidelines available for either researchers or practitioners. This is a serious obstacle for educators who want to improve metacognition in the classroom. One difficulty is understanding what metacognition is, a problem that is due in part to too much theoretical and terminological slippage. Another problem is selecting a benchmark for measuring the growth of metacognition (see Suggestion 3 above). A third problem is deciding on a method for assessing metacognitive progress. My observations have led me to conclude that teachers and parents typically assess metacognitive growth using intuitive hunches rather than tangible outcome measures. This is due to at least three factors: (a) uncertainty about what to look for; (b) the lack of meaningful, cost-efficient measurement strategies; and (c) the lack of meaningful interpretative guidelines.

The lack of procedural and interpretative guidelines is a serious problem within and beyond the field of metacognition. Within the field, there are few established procedures for measuring metacognitive knowledge. Of the three most widely used paper-and-pencil instruments (i.e., the LASSI, MSLQ, and MAI), only the LASSI has been used enough to provide relatively stable norms for different types of students. Most studies continue to use instruments designed specifically for the study at hand. Many studies that used the LASSI, MSLQ, or MAI also have fairly small sample sizes (i.e., less than 100), that preclude a meaningful replication of previous studies. To complicate matters, very few studies ever report reliability coefficients for criterion measures, and few researchers have enough measurement and statistical savvy to understand the subtle, yet important, differences among alternative data-analytic strategies once data have been collected (Keren, 1991; Schraw, 1995).

Another issue concerns the purpose of metacognitive assessment (Baker & Cerro, this volume). Information gathered during an assessment can be used for different purposes (Tindal & Marston, 1990). One purpose is to make placement decisions, such as whether a
student is admitted to a restricted program. A second purpose is to make diagnostic judgments regarding a specific skill or learning disability. A third purpose is to use the information to provide ongoing feedback to students, parents, and teachers. A fourth purpose is to provide summative evaluation at the end of an instructional unit or training session. Whereas most researchers used measures of metacognition as summative indices of unobservable metacognitive competencies, most educators are interested in metacognition from a diagnostic and formative evaluation perspective. There has been very little thought given to bridging the gap between researchers and theorists on the one hand, and practitioners on the other. I believe researchers have a responsibility to bridge this gap and to address how educators can collect information about a student’s skills in a manner that enables the educator to provide useful formative feedback. Similarly, many teachers hunger for guidelines for diagnosing potential learning difficulties related to lack of cognitive and metacognitive regulation.

Suggestion 6: Establish guidelines for implementing and evaluating instruction. Metacognitive theory has not been translated adequately into educational practice. This sentiment was captured well by Baker (1989), Garner (1987), and Baker and Cerro (this volume). Few teachers have a clear sense of how to improve metacognition and metamemory, or even ways to enhance the growth of specific subcomponents of metacognition such as conditional knowledge, although some educators are quite skilled in this regard. Educators desire guidelines for helping their students become more metacognitively aware. These guidelines should explicate separate subcomponents of metacognition (e.g., conditional knowledge, monitoring) as specifically as possible, and propose specific instructional interventions that improve these skills. Researchers also must propose guidelines for assessing the growth of metacognitive skills.

A number of researchers have investigated the kind of instructional interventions I am describing (Brown & Pressley, 1994; Delclos & Harrington, 1991; King, 1991; Paris & Jacobs, 1987; Palincsar & Brown, 1989; Zimmerman & Martinez-Pons, 1990). Many of these interventions have focused on teaching specific strategies (e.g., identifying main ideas) and monitoring their use. Others have proposed a broader research agenda (Garner, 1990; Pressley, Harris, & Marks, 1992; Van Meter, Yokoi, & Pressley, 1994). My own view is that instructional research in metacognition should proceed in several ways. One is the traditional theory-driven approach reflected in the work of Palincsar and Brown (1984). Another avenue is to utilize in-
class observations to construct grounded theories of metacognition (Strauss & Corbin, 1990; Pressley, this volume). Grounded theories seem especially important at this juncture given the lack of a unified theory of metacognition. A third approach is to utilize phenomenological methods (Creswell, 1994; Moustakas, 1994) that provide an in-depth descriptive account of what effective teachers do to improve metacognition.

Educators need to assess the effectiveness of their interventions as well. As I have suggested at several points in this chapter, guidelines and proven instruments for doing so are lacking. There is much that teacher-preparation programs could do to enhance future educators' knowledge of measurement and assessment. However, at present, many preservice teachers appear to be ill-prepared to meet assessment challenges. At a minimum, classes in reading, science, and mathematics instruction should include methods for providing a knowledge base, as well as metacognitive knowledge about regulating that knowledge base. Suggestions for assessing the effectiveness of both kinds of instruction (i.e., cognitive and metacognitive skills) should be embedded within this context.

Suggestion 7: Consider the relationship among metacognitive and affective variables. This suggestion follows from the work of Borkowski and colleagues (Borkowski & Muthukrishna, 1992; Borkowski, Millstead, & Hale, 1988; Borkowski et al., this volume) that describes the relationship among metacognitive and affective variables. Far less has been made of these important connections than they deserve (see Weinert, 1987). However, several studies suggest that metacognition may play a role in increasing personal interest and reducing anxiety that interferes with task performance (Tobias, 1995; Tobias & Everson, this volume). Similarly, metacognition may facilitate the understanding and regulation of emotions and possible selves in academic settings (Borkowski & Thorpe, 1994). Last, Pintrich and colleagues (Garcia & Pintrich, 1994; Pintrich & DeGroot, 1990; Pintrich et al., this volume) have elaborated on a number of important connections between metacognition and motivation.

Summary

This section has proposed a broad agenda for translating metacognitive theory into educational practice on a broad scale. This goal is ambitious. It will take the better part of a decade under the best of circumstances. The success of this research depends in large part on the ability of theorists, instructional design experts, and
specialists in educational measurement to carve out a common agenda. This agenda should accomplish the following: (a) provide operational definitions of specific metacognitive skills and their relationship to the student's knowledge base, (b) construct and evaluate a variety of outcome measures that can be used to assess these skills, (c) cross-validate these outcome measures so that they can be used collectively to conduct multidimensional assessment of metacognitive skills, and (d) train practicing teachers to teach and evaluate metacognitive skills.

Conclusion

This chapter has attempted to provide an overview of salient measurement issues that are relevant to the study of metacognition. Although I believe the construct of metacognition is essential to understanding human cognition, the past two decades of research and practice have not achieved the lofty goal of presenting a comprehensive theory of metacognition that can be rendered into educational practice. It is time to take this goal seriously. The Buros Symposium on Issues in the Measurement of Metacognition was convened for this specific purpose. The chapters included in this volume have turned over many rocks, showing us the good, the bad, and the ugly.

Yet although a call to arms is warranted, it would be misleading to suggest that the field has reached an impasse. As evidenced by the chapters in this book, researchers agree on many essential points regarding metacognitive theory and practice. There is substantial agreement that the related constructs of metacognition and self-regulation have made an enormous contribution to cognitive psychology, literally changing the way that psychologists and educators view cognition and development. There also is widespread agreement that it is time to shore-up our knowledge of metacognition with sound measurement practice.

The separate and combined contributions of this volume point us in the right direction for substantial progress over the next decade. Metacognition has much to offer teachers and researchers. I believe the field of metacognition can deliver on its promise of helping students at all levels understand their thinking and learning.

REFERENCES


7. ASSESSING METACOGNITION


