3. Assessing Metacognition in Children and Adults

Linda Baker
University of Maryland, Baltimore County, baker@umbc.edu

Lorraine C. Cerro
University of Maryland, Baltimore County

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/4

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 in Children and Adults

Linda Baker

University of Maryland, Baltimore County

Lorraine C. Cerro

University of Maryland, Baltimore County

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

Clearly, the construct of metacognition has had wide appeal and wide applicability, stimulating a great deal of research across a broad
spectrum of psychological problems and issues, as well as a growing amount of intervention work in classrooms. In a 1994 review paper on social influences on metacognitive development, Baker wrote, "The popular appeal of metacognitive development has led to the widespread adoption and somewhat uncritical acceptance of the construct among educators. This situation is obviously problematic from a scientific standpoint and makes clear the need for further basic research on how metacognition develops, the role of metacognition in cognitive development, and how metacognition may best be fostered" (pp. 202-203).

The concern about uncritical acceptance is no less apt with regard to measurement; let us therefore amend the final sentence to end with and measured.

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

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

DEFINITIONAL ISSUES

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

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

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

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

METHODS FOR ASSESSING METACOGNITION USED IN BASIC RESEARCH

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

**Verbal Reports**

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

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

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

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

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

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

Error-Detection Approaches

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

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

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

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

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

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

Concerns Expressed about the Measurement of Metacognition

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

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

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

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

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

DEVELOPMENT OF METACOGNITIVE ASSESSMENT TOOLS

Instruments for Assessing Metamemory

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

Instruments Designed to Assess Metacognitive Knowledge in Reading

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

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

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

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

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

Instruments for Assessing Metacognitive Strategy Use in Learning and Studying

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

The instrument developed by Pintrich and his colleagues, the Motivated Strategies for Learning Questionnaire (MSLQ), is a self-report measure of adolescents’ and college students’ motivational orientations and use of various learning strategies (Pintrich & DeGroot, 1990; Pintrich, Smith, Garcia, & McKeachie, 1993). Thirty-one of the items are motivational and 50 are learning strategies. Of these 50, 12 items are concerned with metacognitive self-regulation; specifically tapping the processes of planning, monitoring, and regulating. The instrument has been subjected to extensive checking of reliability and validity. The metacognitive self-regulation scale has an adequate level of internal consistency (.79), and an overall correlation of .30 with final course grade.
3. METACOGNITION IN CHILDREN AND ADULTS

Perhaps the most widely used instrument is the Learning and Study Strategies Inventory (LASSI, Weinstein, 1987), available as a paper-and-pencil test and in a computerized version. The LASSI consists of 77 items rated using a 5-point Likert-type scale that ranges from not at all typical of me (1) to very much typical of me (5). It yields 10 subscale scores: attitude, motivation, time management, anxiety, concentration, information processing, selecting main ideas, study aids, self-testing, and test strategies. The assignment of items to these subscales was based on the intuitive judgment of several experts (Weinstein, Zimmerman, & Palmer, 1988). The LASSI has been subjected to extensive validation efforts, including those by independent researchers (e.g., Olejnik & Nist, 1989), and it is considered to have good psychometric properties.

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

Instruments for Assessing Metacognition in Problem Solving

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

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

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

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

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

Recommendations for Using Interviews

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

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

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

A number of published interview instruments that have been recommended for teachers are now being published in secondary sources, thus giving them what might appear to be even greater legitimacy. For example, Rhodes (1993) published a handbook of informal instruments for assessing literacy that included several metacognitive interviews drawn from other sources. One of the instruments was the 10-question Reading Interview: A reader's view of the reading process (Goodman, Watson, & Burke, 1987). Directions for administering the interview provided by Goodman et al. include coding directions, with categories of responses students might provide. This level of detail seems appropriate and helpful for teachers.

Another interview Rhodes included in her collection was a content reading interview based on Wixson, Bosky, Yochum, and Alvermann (1984). The questions are similar to those used in other interviews, but are tied specifically to a particular content area selected by the interviewer. This instrument has been criticized on the grounds that no reliability data or validity data were provided, but the lack of such data is a common weakness of most of these interview instruments.

Recommendations for using think-aloud measures

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

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

Recommendations for Using Error Detection Procedures

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

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

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

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

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

Recommendations for Assessing Metacognitive Processing in Authentic Tasks

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

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

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

Comments on the Recommendations for Metacognitive Assessments in Classroom and Clinic

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

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

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

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

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

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

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

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

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

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

Portfolio Assessments

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

Commercial Performance Assessments

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

Statewide Performance Assessments

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

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

Sometimes I

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

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

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

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

Additional Recommendations for New Educational Assessments

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

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

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

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

On the Value of Converging Evidence

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

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

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

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

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

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

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

On the Criteria for Evaluating Metacognitive Assessments

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

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

On the Uses to Which Tests of Metacognition are Put

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

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

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

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

Summary

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

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

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

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

REFERENCES


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

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

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

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

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


New York: Addison-Wesley.

York: The Guilford Press.

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

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


a scale used to measure metacognitive reading awareness. Journal of Educational Research, 85, 81-87.


Meltzer, L. (1991). Problem-solving strategies and academic per-
formance in learning-disabled students: Do subtypes exist? In L. V.
Feagans, E. J. Short, & L. J. Meltzer (Eds.), Subtypes of learning


