Reliability and Validity of a State Metacognitive Inventory: Potential for Alternative Assessment

CSE Technical Report 469

Harold F. O'Neil, Jr., Project Director
CRESST/University of Southern California

Jamal Abedi
CRESST/University of California, Los Angeles

December 1996

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported in part under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education and in part under National Center for Education Statistics Contract No. RS90159001 as administered by the U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the policies of the Office of Educational Research and Improvement, the National Center for Education Statistics or the U.S. Department of Education.

RELIABILITY AND VALIDITY OF A STATE METACOGNITIVE INVENTORY: POTENTIAL FOR ALTERNATIVE ASSESSMENT

Harold F. O'Neil, Jr.
CRESST/University of Southern California

Jamal Abedi
CRESST/University of California, Los Angeles

Abstract

An assumed advantage of alternative assessments is that they result in more higher level thinking or metacognitive skills. We believe that this advantage should be measured directly and explicitly. Unfortunately, few standardized measures of metacognitive skills (planning, monitoring, cognitive strategies, and awareness) exist. In our studies, for 12th graders, alpha reliability estimates and factor analysis indicated that our metacognitive subscales are reliable (alpha above .70) and unidimensional (one factor per subscale). Because the subscales have only 5 items each, they meet brevity standards. Construct validity of our state metacognitive inventory is acceptable. Results indicate that our state metacognitive inventory yields useful information about both the assessment and students.

Alternative assessments share several common characteristics. Herman, Aschbacher, and Winters (1992) provide an excellent listing of such characteristics: (a) Ask students to perform, create, produce, or do something; (b) tap higher level thinking and problem-solving skills; (c) use tasks that represent meaningful instructional activities; (d) invoke real-world applications; (e) people, not machines, do the scoring, using human judgment; and (f) require new instructional and assessment roles for teachers (p. 6).

Performance assessment is by nature a process that requires extended engagement by students in order to demonstrate their proficiency. They may conduct multistep experiments, write well-documented research papers, organize and supervise group problem solving, or present a description of previously developed work. Although the exact nature of these tasks may differ

in terms of subject matter, time for performance, flexibility or choice of topics, and the amount of external support for the student, they share the common characteristic of requiring that students plan, organize and execute complex tasks.

When inspecting students' performance results on these assessments, which, for the most part, have been relatively disappointing (O'Neil & Brown, 1995), a series of alternative hypotheses arise. Perhaps students do not perform well because they have not been taught the material; perhaps their low level of performance relates to their lack of prior relevant knowledge; or perhaps they have not learned how to structure and manage their time well in the accomplishment of the target tasks. Our work was undertaken to produce information about the collateral skills necessary to the accomplishment of complex performance—a student's ability to think about the task systematically. The measure is also intended to be a useful indicator for those educational goals that emphasize work habits or metacognitive strategies.

There are many assumed advantages of alternative assessments, for example, that such assessments should result in more effort expended and perhaps less anxiety. Further, such assessments should engage students in more higher level thinking or metacognitive skills. We believe there is a need to measure such assumed advantages directly and explicitly. Unfortunately, few standardized or commercially available measures of effort, anxiety or metacognitive skills exist.

As part of an ongoing CRESST R&D effort in developing new measures for alternative assessment, we have been designing, developing, and validating a set of self-regulation measures for use with such alternative assessments. We view self-regulation as consisting of the constructs of metacognition, effort, and anxiety. This paper will address the reliability and validity of our newly developed measure of metacognition. We view metacognition as consisting of planning, monitoring, cognitive strategies and awareness. The measure has been validated in a series of experimental studies (Khabiri, 1993; Kosmicki, 1993; O'Neil, Sugrue, Abedi, Baker, & Golan, 1992; Yap, 1993).

Our framework for test development in metacognition is domain-independent assessment. Domain-independence is independent of domain (task, subject matter) but tied to either a type of learning (e.g., metacognitive) or affect

(e.g., anxiety). However, a domain-independent measure must be instantiated in a context (e.g., assessment, learning task).

Our concept of metacognition is derived from that of Pintrich and DeGroot (1990). They suggested that metacognition consists of strategies for planning, monitoring and modifying one's cognitions. We also view metacognition as composed of planning, monitoring or self-checking, and cognitive strategies. We have added the construct of awareness as we believe there is no metacognition without being consciously aware of it (see also Flavell, 1979). Further, in contrast to existing measures of metacognition, we view these constructs from both a cognitive science perspective (e.g., Barsalou, 1992; Beyer, 1988; Hayes-Roth, 1988) and a state-trait perspective (e.g., Spielberger, 1975). Finally, we have been informed by the other research on the measurement of metacognition (Borkowski & Muthukrishna, 1992; Everson, Smodlaka & Tobias, 1994; Paris, Cross, & Lipson, 1984; Pintrich & DeGroot, 1990; Pressley & Afflerbach, 1995; Tobias & Everson, 1995; Zimmerman, 1989; Zimmerman & Martinez-Pons, 1986, 1990, 1988).

State-Trait Conceptions

Using constructs from state-trait anxiety theory (Spielberger, 1975) as an analogy, we have formulated a set of self-report, domain-independent trait and state measures of metacognition. We find the state versus trait distinction useful for both cognitive and affective measurement. Thus, we have generalized the key constructs from an affective domain (e.g., state and trait anxiety) to a cognitive domain (i.e., state and trait metacognition).

States are situation-specific and are considered to vary in intensity and change rapidly over time. We define state metacognition as a transitory state of people in intellectual situations, which varies in intensity, changes over time, and is characterized by planning, monitoring or self-checking, cognitive/affective strategies, and self-awareness. Traits are considered relatively enduring predispositions or characteristics of people (e.g., intelligence or aptitude). We define trait metacognition as a relatively stable individual difference variable to respond to intellectual situations with varying degrees of state metacognition. In this paper we will discuss only state metacognition.

In summary, we define metacognition as the conscious and periodic self-checking of whether one's goal is achieved and, when necessary, selecting and

applying different strategies. One is self-aware of the process in the following ways. Planning: One must have a goal (either assigned or self-directed) and a plan to achieve the goal. Self-monitoring: One needs a self-checking mechanism to monitor goal achievement. Cognitive strategy: One must have a cognitive or affective strategy to monitor either domain-independent or domain-dependent intellectual activity (for example, finding the main idea is a domain-dependent cognitive strategy). Awareness: The process is conscious to the individual.

The following items are examples of state metacognitive items. Planning: *I tried to understand the task before I attempted to solve it*; Self-checking: *I checked my work while I was doing it*; Cognitive strategy: *I used multiple thinking techniques or strategies to solve the task*; Awareness: *I was aware of my ongoing thinking processes*.

The techniques for measuring metacognition in empirical studies may be categorized into two kinds: domain-dependent and domain-independent. One of the major domain-dependent methodologies is think-aloud protocol analysis. In this technique, a subject is asked to vocalize his or her thinking processes while working on a problem. The data as a protocol are then coded according to a specified model for psychological analysis, which provides insights into elements, patterns, and sequencing of underlying thought processes. An excellent review of mainly domain-dependent metacognitive assessment techniques including protocol analysis is provided by Royer, Cisero, and Carlo (1993). Another interesting domain-dependent technique in reading is provided by Everson et al. (1994).

There are several interesting domain-independent measures of cognitive and affective processes (see, for example, Pintrich & DeGroot, 1990; Weinstein, Palmer, & Schultz, 1987) to measure metacognition. These investigators use rating scales to measure metacognition. This type of measurement involves asking participants to answer or self-report on statements about cognitive or affective processes. For example, to measure learning strategies, a commercially available self-rating inventory is *The Learning and Study Strategies Inventory (LASSI)* (Weinstein et al., 1987). This self-report inventory measures learning and study strategies, for example, (a) attitude and interest; (b) use of time management principles for academic tasks; (c) anxiety and worry about school performance; (d) information processing, acquiring knowledge, and reasoning, and (e) test strategies and preparing for tests. However, this inventory was

conceptualized and developed before much of the current research on metacognition and reflects an eclectic view of both cognitive and affective processes. According to our definition of metacognition, the *LASSI* does not measure metacognition. Another interesting self-rating scale on motivational beliefs and self-regulated learning, the Motivational Strategies for Learning Questionnaire (MSLQ) (Pintrich & DeGroot, 1990), does not explicitly address either the state-trait distinction or specific metacognitive constructs (e.g., planning), which we believe are critical in the measurement of metacognition.

Reliability and Validity of State Measures

As with our use of state and trait constructs from state-trait anxiety theory (Spielberger, 1975; 1983) to define state metacognition, our approach to reliability for our state metacognitive measure is also based on an analogy from state-trait anxiety theory. Spielberger (1972) discussed three important requirements of state anxiety measures: brevity, reliability, and ability to reflect stress. With respect to state metacognition, we view the three similar requirements to be brevity, reliability, and ability to reflect varying intellectual demands of tasks or tests. Spielberger (1972) recommended brevity for state measures because long, involved scales would be unsuitable for experimental tasks in which administration of an extensive measure could interfere with performance on the task.

Spielberger recommended internal consistency as the type of reliability suitable for state anxiety measures because anxiety states vary in intensity and fluctuate over time. Thus, it was assumed and later demonstrated that test-retest correlations would be nonsignificant. Only in the case in which a person is placed in the same situation on retest would one expect a high degree of relationship between two state anxiety measures taken at two different times. This expectation was also confirmed by O'Neil (1972).

Because our state metacognitive inventory was to be employed in the context of assessment and learning tasks, it was feared that a scale with many items could, by its length, interfere with performance on the task itself. It was hoped that a briefer state measure, which could be administered in less than a minute, would still provide reliable and valid information. Our past experience with brief state anxiety scales (O'Neil, Baker, & Matsuura, 1992; O'Neil & Richardson, 1977) indicated that a scale of 5 items per measure might meet those

requirements for brevity and good internal consistency. Thus, the entire state metacognitive inventory was designed to be 20 items, with 5 items for each of four subscales (planning, monitoring, cognitive strategies and awareness).

Our approach to validation relies heavily on construct validity techniques as well as content validity. Consistent with a construct validity approach, we make the following predictions regarding state metacognition: (a) Planning, self-checking, cognitive strategies, and awareness would be positively related; (b) state metacognition would be more predictive of achievement than trait metacognition; (c) higher levels of state metacognition would lead to better academic performance; (d) higher levels of state metacognition would be exhibited on more difficult tasks; (e) persons with higher education levels would exhibit higher levels of state metacognition.

Early Development

Our work on state metacognition inventory construction began with measurement issues identified in our research on the human benchmarking of expert systems (O'Neil, Baker, Jacoby, Ni, & Wittrock, 1990; O'Neil, Baker, Ni, Jacoby, & Swigger, 1994). An expert system is computer software that can accomplish a task that a human expert can. Human benchmarking is an evaluation procedure by which an expert system's performance is judged based on a sample of people's performance (both on processes and outcomes) on tasks with psychological fidelity. It is a variation of the Turing Test (Turing, 1988). The context of our human benchmarking research was the expert system GATES. The GATES program schedules airplanes to gates by assigning an airplane to a specific gate, time, etc., without violating constraints.

GATES is an expert system written in Prolog for gate assignment of airplanes at TWA's JFK and St. Louis airports (Brazile & Swigger, 1988, 1989). We considered the software processes or rules in GATES to be like or analogous to state metacognition in people. We developed a problem-solving task (based on the GATES program) that requires people to solve the same task as the program. For this scheduling task, the expert system and people have the same goal: to assign all landed flights to available gates. Both the system and people are assumed to follow the same constraints and rules to do the task. Following these restrictions, the expert system monitors itself in three phases of scheduling. People may use the same constraints to plan, monitor, and assess their ongoing

processes of scheduling. However, people are aware of their ongoing metacognitive processes while the expert system is not. Thus, the psychological process equivalent of GATES scheduling is metacognition.

To use our human benchmarking paradigm we needed to measure this common process of metacognition (with GATES) in people. When we found no such measure in the literature, we began the development process to create one. The state self-monitoring questionnaire's goal was to determine whether students were engaged in metacognition while doing the scheduling task. The original 26 items asked about students' planning, monitoring, cognitive strategy use, and awareness. For example, a sample item for planning was "I explicitly planned my course of action," to which a student answered: 1—*not at all*; 2—*somewhat*; 3—*moderately so*; or 4—*very much so*. A single, total metacognition score was generated.

In this initial study individuals with different educational levels (with assumed different ability and metacognitive levels) performed as predicted; that is, university undergraduate students displayed significantly more state metacognitive activity than community college students (see Table 1). University undergraduate students performed significantly better on the GATES task than community college students. Reliability indices were in the low .90s, and the correlational relationship between state metacognition and performance was .46. These findings also support the construct validity for the measure.

Table 1

State Metacognition on the Scheduling Task

Variable	Community college students			University undergraduates		
	M	*SD*	*N*	*M*	*SD*	*N*
State metacognition	73.01	14.14	21	82.46	10.61	27
Total correct state assignments	11.96	2.09	25	13.30	1.33	27

Note. Alpha reliability for the university sample was .91 ($N = 99$). Alpha reliability for the junior college sample was also .91 ($N = 21$).

In summary, the self-monitoring measure looked very promising. However, we had a general measure of state metacognition, not subscales for

planning, monitoring, cognitive strategy and awareness. Thus, more items needed to be written and validated to develop the four subscales (planning, self-checking, cognitive strategy, and awareness). The following section presents the results of this developmental effort.

Development of Subscales for the State Metacognition Inventory

New items were written to be consistent with our constructs of planning, self-checking, cognitive strategies and awareness and labeled as the state metacognitive inventory. The state metacognitive inventory was administered to multiple groups of students in successive studies to examine its psychometric characteristics. A common statistical methodology was employed. Descriptive statistics such as means and standard deviations were obtained for each item and each subscale. A classical measure of reliability, Cronbach's alpha, was also obtained to examine internal consistency for the items within each subscale. An item-remainder correlation of each item with the subscale score was also computed. The item-remainder correlation identified how well an item fit within a particular subscale. To further evaluate the internal consistency of items, a principal components factor analysis with varimax rotation was also performed on the items within each subscale to see if any of the subscales was multidimensional.

A set of mathematics achievement tests were used as criterion measures to determine the construct validity relationship between achievement and the various aspects of state metacognition. Based on the descriptive statistics, internal consistency measures, and the results of factor analysis, poor items were identified and removed, and the number of items was reduced. Items were eliminated so that there was no significant reduction in the reliability or validity indices of the subscales.

The following sections of this paper summarize the analyses performed on the state metacognitive inventory. We will report the results in three different sections, which represent three different sets of empirical studies. Each study was conducted with the multiple objectives of (a) investigating the improvement of math achievement by using various experimental treatments, and (b) collecting information on the reliability and validity of the state metacognitive inventory. Reporting this latter objective is the purpose of this paper. The reader is referred to the primary sources for the results of the other objective.

Community College Sample

The state metacognitive inventory consisted of four subscales of metacognition: planning, self-checking, cognitive strategy, and awareness. The entire inventory was administered to a group of 219 community college students along with a 20-item math test (Kosmicki, 1993). The purpose of this study was twofold: (a) to examine the relationship of metacognitive processes and math performance, and (b) to determine the impact of experimentally-manipulated testing conditions (e.g., use of different types of motivational test instructions) on community college students' math performance. The overall hypothesis of this study was that subjects exposed to appropriate instructions before taking a standardized test will demonstrate higher performance and will also produce higher levels of metacognitive processing. However, in this study, there were no significant effects on achievement due to the differential test instructions.

Table 2 presents the number of items and Cronbach's alpha coefficients for the subscales of the state metacognitive inventory for both the full and reduced versions. (The full version refers to the initial set of items that were used in data collection; the reduced version refers to the final set following statistical analyses and revision.) The reliability levels for the full state subscales were acceptable and ranged from .77 for self-checking to .81 for cognitive strategy.

However, there were too many items to meet the brevity criterion for a state measure. Thus, a set of analyses were conducted to reduce the number of items in each subscale. Analyses were done on individual items within each subscale to see how items performed. We will describe in some detail this first iteration of the revision of the state metacognitive inventory. The remaining two sets of empirical studies follow the same line of logic.

We compared item means, item-remainder correlations, factor loadings, commonalties, and reliability coefficients. For the state awareness subscale, item means ranged from 2.70 to 3.15. Item-remainder correlations ranged from .33 to .58. The alpha coefficient of 8 items for this subscale was .78. Based on the summary results of the analyses done on items in the awareness subscale, 2 items were omitted because they had relatively low item-remainder correlations (.33 and .34 respectively), and both of them had moderate loadings on a second factor. Thus, in the reduced version of the instrument, the awareness subscale had only 6 items with an alpha reliability of .79 (see Table 2).

Table 2

Number of Items, Number of Factors, and Alpha Coefficients for the Full and the Reduced State Metacognitive Inventory for the Community College Sample

	Number of items		Number of factors		Alpha	
Subscale	Full	Reduced	Full	Reduced	Full	Reduced
Awareness	8	6	2	1	.78	.79
Cognitive strategy	14	8	4	1	.81	.81
Planning	9	5	2	1	.80	.83
Self-checking	8	5	2	1	.77	.75

Similarly, the results of analysis for the cognitive strategy subscale (14 items) indicated that the item means ranged from 2.00 to 3.30. Alpha reliability for this subscale of 14 items was .81. The items in this subscale loaded on four factors, indicating that all the items within the subscale did not belong to the same category. By looking at the percent of variance extracted by each factor, however, it was noted that most of the items had high loadings on the first factor. The percent of variance extracted by the first factor was 31.6% as compared with 10.0%, 8.1%, and 7.3% for the second, third and fourth factors respectively. Based on the results of analyses done on items within this subscale, 6 items were removed: 1 item because of low item-remainder correlation (.39), a low factor loading, and low communality; and 2 items because of their loading on the third factor. These two items mainly created Factor 3 for this subscale. Removal of these 2 items eliminated Factor 3 and created a more homogeneous set of items under the subscale. Another item was removed because of its negative item-remainder correlation. And 1 item was removed because it was very similar in content to another item. The reduced cognitive strategy subscale had 8 items with an alpha reliability of .81 (see Table 2).

The results of analyses for the planning subscale with 9 items indicated that the item means ranged from 2.13 to 3.22. Item-remainder correlations for this subscale ranged from .17 to .62. The 9 items of this subscale loaded on two factors, Factor 1 explaining 41.3% of the variance and Factor 2, 14.4% of the variance. The alpha coefficient for this 9-item subscale was .80. The results of the analyses performed on the items suggested the omission of the following: 1 item because of relatively low item-remainder correlation (.38) and a non-significant loading

on the first factor; another item because of low item-remainder correlation (.17); a third item was dropped because of higher loading on a second factor. The reduced planning subscale had 5 items with an alpha reliability of .83 (Table 2).

The results of analyses for the state self-checking subscale indicated that the alpha coefficient for this subscale with 8 items was .77. The item means ranged from 2.65 to 2.89. Item-remainder correlations ranged from .38 to .64. Six items loaded on the first factor and two loaded on the second factor. These two items, which also had relatively lower item-remainder correlations with the total subscale, were removed in order to increase internal consistency of the items. A third item was removed that had a low item remainder correlation. The reduced self-checking subscale of 5 items had an alpha reliability of .75 (Table 2).

In summary, we removed 15 items from the different subscales. As Table 2 shows, removing poor items in most cases increased the reliability of the subscales and reduced the number of items to a more manageable level. There were originally 39 items in the inventory. From the total items, 15 items (about 38% of the original items) were removed, yet the reliabilities remained about the same. Another point regarding the reduced versus the full set of items is the reduction in number of factors in the reduced set of items (see Table 2). Principal components factor analysis with varimax rotation analyses yielded either two or four factors for the subscales of the full form; that is, the subscales in the full form were not unidimensional. The problem of multidimensionality of items in the full form created difficulties conceptually and when computing subscale scores. Items under all subscales loaded on only one factor in the reduced form.

After identifying and removing the poor items, the resulting inventory had more homogeneous items within the subscales with acceptable reliability and was quicker to administer. However, we achieved this result with multiple analyses of the same data set. Thus, we decided to use this revised state metacognitive inventory on another group of younger subjects to examine the psychometric properties of the inventory and attempt to replicate the previous findings.

Because our next sample used high school students (Grades 9-12) we expected that some items would behave differently than in our community college sample. Thus, 5 new items were added to the planning subscale, and 3 new items were added to the self-checking subscale. As a result of these changes,

a 32-item inventory resulted. This revised state metacognitive inventory was administered to a group of 230 high school students (Khabiri, 1993).

Initial High School Sample (Grades 9-12)

One purpose of this study (Khabiri, 1993) was to provide reliability and validity information on the subscales of the revised state metacognitive inventory. Another purpose was to test the differential validity among these constructs and their relationships to math performance. It was hypothesized that certain cognitive processes needed for successful math performance would be differentially predicted by planning, self-checking, cognitive strategy, and awareness. However, Khabiri's (1993) study indicated that the differential validity of the subscales was weak.

As in the community college study, we analyzed the full set of 32 items and produced a reduced set. Means and standard deviations as well as alpha coefficients for each of the subscales were computed, and principal components factor analysis with varimax rotation was applied on the subscale items to see how items grouped together under each subscale. Table 3 reports number of items, number of factors, and alpha coefficient for each of the four subscales for the full set of items and the reduced set of items. Alpha reliability coefficients were approximately the same after item deletion. However, the alpha coefficients for the awareness and cognitive strategy subscales were around .70, a minimally acceptable level. Further, multiple factors emerged on two of the subscales (planning and cognitive strategy).

Table 3

Number of Items, Number of Factors, and Alpha Coefficients for the Full and the Reduced Revised State Metacognitive Inventory for the Initial High School Sample

	Number of items		Number of factors		Alpha	
Subscale	Full	Reduced	Full	Reduced	Full	Reduced
Awareness	6	5	1	1	.70	.71
Cognitive strategy	8	7	2	2	.71	.71
Planning	10	9	2	2	.81	.81
Self-checking	8	7	1	1	.79	.75

Table 4 compares the community college state metacognitive inventory (24 items) with the 28-item reduced state metacognitive inventory, initial high school version. Note that an additional 5 new planning items and 3 new self-checking items were added to the item pool. Item deletion resulted in two of the subscales (awareness, cognitive strategy) having slightly lower the alpha coefficients.

Table 4

Number of Items, Number of Factors, and Alpha Coefficients for the Community College Sample and the Initial High School Sample

	Number of items		Number of factors		Alpha	
Subscale	Comm. college	High school	Comm. college	High school	Comm. college	High school
Awareness	6	5	1	1	.79	.71
Cognitive strategy	8	7	1	2	.81	.71
Planning	5	9	1	2	.83	.81
Self-checking	5	7	1	1	.75	.79

Note. Comm. college = Community college.

The comparison of the 24-item community college inventory with the high school inventory may not be valid because the statistics are based on two different groups of subjects (community college students vs. high school students), which may represent two different populations. Thus, any difference in the size of alpha may be attributable to initial differences between the two groups. However, because very similar results were obtained on the subscales with about the same number of items in the full and the reduced forms, we believe the two groups of subjects may be considered to be drawn from the same population.

The National Assessment of Educational Progress (NAEP) Studies

As was true in the prior studies, this set of studies investigated two objectives: (a) the impact of various experimental treatments on test performance, and (b) the reliability and validity of the state metacognitive inventory. One of the major validity questions that has been raised in relation to the National Assessment of Educational Progress (NAEP) concerns the possible

impact of motivational factors on NAEP results. If students are not motivated to perform well on NAEP tests, and if the lack of motivation results in poor performance, then NAEP findings are underestimates of student achievement.

To test the theory that increased motivation to perform well on NAEP would be reflected in increased effort and improved performance on the test, a series of studies was conducted by UCLA's Center for the Study of Evaluation (CSE) and its National Center for Research on Evaluation, Standards, and Student Testing (CRESST). The studies investigated the effects of various motivational conditions on the performance of 8th- and 12th-grade students on a subset of released items from the 1990 NAEP mathematics test.

In order to link any observed performance differences to differential investment of effort or to differences in metacognition or anxiety, these variables were measured via a modified self-assessment questionnaire (O'Neil, Sugrue, et al., 1992). A number of pilot studies were conducted to select the motivational conditions that might influence test performance.

12th-Grade Financial Incentive Pilot Sample

One pilot study compared the performance of 12th-grade students who received three different financial rewards (e.g., 50¢ per test item correct). The study yielded no differences among test scores of 12th-grade students who received any of the three financial incentives versus students who received standard NAEP test instructions.

A modified version of the state metacognitive inventory with four subscales was administered to the 12th-grade students in the pilot study. The number of items in the self-checking subscale was increased from 7 to 11. Table 5 summarizes the descriptive statistics for the subscales used in the 12th-grade pilot study compared to the initial high school sample (Khabiri, 1993). As Table 5 indicates, subscale reliabilities ranged from .82 for awareness to .87 for self-checking.

Factor analyses were performed on items within the subscales. The results of these analyses are also shown in Table 5. The cognitive strategy and planning subscales have two factors each. As the same set of items was used, this was not unexpected. Table 5 also compares subscale statistics between the initial high

school sample (Grades 9-12) and the 12th-grade sample. Alpha reliability increased for all subscales.

Table 5

Number of Items, Number of Factors and Alpha Coefficients for Initial High School Sample and the 12th-Grade Sample

Subscale	Number of items		Number of factors		Alpha	
	High school	12th grade	High school	12th grade	High school	12th grade
Awareness	5	5	1	1	.71	.82
Cognitive strategy	7	7	2	2	.71	.83
Planning	9	9	2	2	.81	.84
Self-checking	7	11	1	1	.79	.87

8th-Grade Sample (Main Study)

Based on previous research and on our feeling that 50¢ per item might not be enough to motivate Los Angeles teenagers, a financial incentive condition offering a larger reward of $1 per correct item was included in the main study. The main study compared the effects of three experimental motivational conditions (financial reward, competition, personal accomplishment) and standard NAEP test instructions on the mathematics performance. The results indicated that the offer of a financial reward can improve the performance of 8th-grade students (O'Neil, Sugrue, et al., 1992; O'Neil & Sugrue, in press).

Because some of the pilot study students could not answer all the metacognitive questions, we decided to reduce the number of items even further based on the pilot study results and based on the National Center for Education Statistics (NCES) staff input on item sensitivity. We reduced the number of items in each of the subscales to 5. Therefore, two different versions of the inventory were prepared. For the 8th-grade students, a version with two subscales was used: cognitive strategy and self-checking. For 12th-grade students all four subscales of the inventory were used. Over 95% of both 8th- and 12th-grade students in the main study administration answered all the metacognitive questions.

Table 6 summarizes the results of analyses of the two subscales for the 8th-grade students in the main study. Alpha coefficients for 8th-grade students were low. The alpha coefficient for cognitive strategy was .61, and for self-checking was .64. The low reliability of the subscales for the 8th-grade students may be due to difficulty in the vocabulary of the items However, we achieved our desire of having only one factor per subscale.

Table 6

Number of Items, Mean, Standard Deviation and Cronbach's Alpha for the Main Study, 8th Grade

Variable	# of items	*M*	*SD*	# of factors	Alpha
Cognitive strategy	5	2.75	.65	1	.61
Self-checking	5	2.68	.63	1	.64

12th-Grade Sample (Main Study)

A financial incentive condition offering a larger reward of $1 per item correct was included in the main study. The main study compared the effects of three experimental motivational conditions (financial reward, competition, personal accomplishment) and standard NAEP test instructions on the mathematics performance of 12th-grade students. For 12th-grade students, a fifth condition was added: Students were offered a certificate of accomplishment if they scored in the top 10% of their class. There was no impact of any incentive on 12th grade students (O'Neil, Sugrue, et al., 1992; O'Neil & Sugrue, in press).

The results of the analyses done at the item level for each subscale for the 12th-grade students are summarized in Table 7. Subscale means ranged from 2.52 for self-checking to 2.84 for awareness. These results are very similar to the results obtained for 8th-grade students, but the subscale reliabilities for the 12th-grade students were higher than those for the 8th-grade students. The alpha coefficients of the four subscales for 12th-grade students ranged from .73 for self-checking to .78 for awareness and planning. We factor analyzed item-level data for the 12th-grade subjects of the main study. The results indicated only one factor per subscale.

Table 7

Number of Items, Mean, Standard Deviation and Cronbach's Alpha for the Main Study, 12th Grade

Variable	# of items	*M*	*SD*	# of factors	Alpha
Awareness	5	2.84	.70	1	.78
Cognitive strategy	5	2.66	.73	1	.77
Planning	5	2.76	.72	1	.78
Self-checking	5	2.52	.68	1	.73

Finally, for 12th graders, Table 8 compares the last reduced version of the instrument (20 items) with the prior 12th-grade version (33 items, pilot study). We compare the versions in number of items, number of factors and the size of alpha. As Table 8 indicates, the number of items was reduced from 33 to 20.

Table 8

Number of Items, Number of Factors and Alpha Coefficients for the 12th-grade Pilot Study and the Reduced 12th Grade Main Study

	Number of items		Number of factors		Alpha	
Subscale	Full	Reduced	Full	Reduced	Full	Reduced
Awareness	5	5	1	1	.82	.78
Cognitive strategy	8	5	2	1	.83	.77
Planning	9	5	2	1	.84	.78
Self-checking	11	5	1	1	.87	.73

The self-checking subscale was reduced most dramatically, from 11 items to 5 items. The subscales in the initial version had either one or two factors. In the final-version inventory, however, all items within any of the four subscales loaded on only one factor, which means that under each category there was only one category on one dimension of items, and we had more homogeneous sets of items under the revised subscales than in the original form. The alpha coefficients of the subscales of the original and the final versions were close with the exception of self-checking. Reduction of items did have some effect on the

reliabilities of the subscales: Reliabilities of the final subscales were in the .70s while those of the longer version were in the .80s.

As indicated earlier, comparing the original form with the reduced form on two different groups of subjects may not be a valid comparison; however, comparable results of the two forms obtained from two different groups indicate that, in a sense, the subscales were cross-validated. As mentioned earlier, principal components analysis was performed on the items within each subscale to see if items were unidimensional within a subscale. Normally, a confirmatory factor analysis should follow exploratory analysis to see if the selected items fit under a specific subscale. Confirmatory factor analysis, however, was not done because of the limitation of number of subjects within any single study group. Combining different groups of subjects who were given the metacognitive instrument could result in enough subjects to satisfy the confirmatory analysis subject requirement, but the problem in combining the groups is the lack of exact comparability of metacognitive items across the groups of subjects and various experimental treatments in each study. An additional study in our lab with an appropriate number of subjects using confirmatory factor analysis supported our hypotheses with respect to dimensionality (Yap, 1993).

Relationship of State Metacognition With Achievement

As was mentioned earlier, our basic design involved investigating the relationship of state metacognition with achievement so as to provide some evidence of construct validity for our state metacognitive inventory. Thus, in each of the prior studies, the relationship of state metacognition with achievement was estimated. These results are shown in Table 9. With the exception of the human benchmarking study, the content of the achievement tests was mathematics. The correlations are in the predicted direction, that is, high state metacognition resulting in high performance. Subsequent research using structural equation modeling (Yap, 1993; Li & O'Neil, 1995) indicates that metacognition was influencing achievement and not vice versa. The correlations are mainly significant but low. The range of such correlations is from .04 to .46 with a median correlation of .18.

Given the acceptable reliabilities of these metacognition measures, the magnitude of these correlations is of concern. A review of the metacognitive literature focusing on this issue alone found few studies that reported the

Table 9

Correlations of State Metacognition With Academic Achievement

Subscale	Human bench-marking (n = 21)	Comm. college (n = 250)	Initial high school (n = 210)	12th-grade pilot	8th-grade main	12th-grade main
Metacognition	.46**	.25**	NA	.35** (n=213)	.18** (n=744)	.23** (n=714)
Awareness	NA	.19**	.03	.33** (n=207)	NM	.22** (n=715)
Cognitive strategy	NA	.17**	.12	.36** (n=213)	.15** (n=745)	.21** (n=715)
Planning	NA	.16**	.10	.30** (n=213)	NM	.17`** (n=715)
Self-checking	NA	.12*	.09	.26** (n=213)	.17** (n=744)	.20** (n=715)

Note. Comm. college = Community college. NA = Not available. NM = Not measured.

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

relationship of metacognition with achievement. Pintrich and DeGroot (1990) reported a range of correlations with various forms of classroom achievement from .07 to .36 with a median correlation of .21. Pintrich (1989) reported correlations of metacognition and various indices of student performance of .31, .29, .19, and .31. Pintrich and Garcia (1991) reported a correlation of .27 with a final grade for college students. Thus, our values, although low, are consistent with the limited literature on the effect of metacognition on achievement. Further, it appears that the relationship is stronger with older students.

Conclusions

In summary, with respect to reliability, we have suggested that the appropriate technique for a state measure was internal consistency. For 12th graders, the results of both alpha reliability estimates and factor analysis indicated that our subscales are reasonably reliable (alpha above .70) and unidimensional (no subscale has more than one factor). Further, since the subscales have only 5 items each, they meet the standard of brevity.

Our major measure of validity was construct validity. With respect to construct validity, the following predictions were preliminarily supported:

(a) planning, self-checking, cognitive strategy, and awareness would be positively related (see Khabiri, 1993; Kosmicki, 1993; O'Neil, Sugrue, et al., 1992); (b) state metacognition would be more predictive of achievement than trait metacognition (see Kosmicki, 1993); (c) higher levels of state metacognition would lead to better academic performance (this paper); (d) higher levels of state metacognition would be required on more difficult tasks (O'Neil et al., 1990); (e) persons with higher education levels would exhibit higher levels of state metacognition (O'Neil et al., 1990). More research is obviously needed.

The findings are most robust for our 12th-grade and older samples. The current version of the 12th-grade state metacognitive inventory (see Appendix) is recommended for research use for measurement of alternative assessments. We also recommend the use of retrospective state instructions, that is, "Tell how you felt during the assessment." Thus, our inventory should be given immediately after an assessment or learning task. Logically, one could thus argue that levels of metacognition are caused by the alternative assessment or that levels of metacognition cause the good/bad assessment scores. Some evidence in our lab indicates that metacognition influences performance and not vice versa (e.g., Yap, 1993).

Since the reliability of the inventory is marginal for 8th graders, the current state metacognitive inventory is not recommended for 8th-graders or younger students. However, we have revised the current 8th-grade version and used it as an assessment of the impact for 8th graders of an alternative assessment (California Learning Assessment System, 1993a, 1993b). We are currently documenting that effort (O'Neil & Brown, 1995). In general, the results indicate that the revised state metacognitive inventory is reliable and yields useful information for 8th graders about both the assessment and students.

In conclusion, our state metacognitive inventory operationally defines students' metacognition as a construct consisting of the following subscales or sub-behaviors: (a) planning, (b) monitoring, (c) cognitive strategy, and (d) awareness. The relationship between the scores of the subscales of this instrument has been investigated, and the results have provided preliminary evidence for the construct validity of this instrument. Thus, metacognition can be directly and explicitly measured in the context of alternative assessments.

References

Barsalou, L. W. (1992). *Cognitive psychology: An overview for cognitive scientists*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Beyer, B. K. (1988). *Developing a thinking skills program*. Boston, MA: Allyn & Bacon.

Borkowski, J. G., & Muthukrishna, N. (1992). Moving metacognition into the classroom: "Working models" and effective strategy teaching. In M. Pressley, K. R. Harris, & J. T. Guthrie (Eds.), *Promoting academic competence and literacy in school* (pp. 477-501). San Diego, CA: Academic Press.

Brazile, R., & Swigger, K. (1988). GATES: An expert system for airlines. *IEEE Expert, 3*, 33-39.

Brazile R., & Swigger, K. (1989). *Extending the GATES scheduler: Generalizing gate assignment heuristics*. Unpublished manuscript.

California State Department of Education (1993a). *Statewide performances: Standards for the California Learning Assessment System (CLAS)* (A supplement to: Students, Standards, and Success). Sacramento, CA: Author.

California State Department of Education (1993b). *High school performance assessment*. Sacramento, CA: CTB Macmillan/McGraw-Hill.

Everson, H. T., Smodlaka, I., & Tobias, S. (1994). Exploring the relationship of test anxiety and metacognition on reading test performance: A cognitive analysis. *Anxiety, Stress, and Coping, 7*, 85-96.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*, 906-911.

Hayes-Roth, B. (1988). A cognitive model of planning. In A. Collins & E. E. Smith (Eds.), *Readings in cognitive science: a perspective from psychology and artificial intelligence*. San Mateo, CA: Morgan Kaufmann Publishers.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Khabiri, P. (1993). *The role of metacognition, effort and worry in math problem solving requiring problem translation*. Unpublished doctoral dissertation, University of Southern California, Los Angeles.

Kosmicki, J. (1993). *The effect of differential test instructions on math achievement, effort, and worry of community college students.* Unpublished doctoral dissertation, University of Southern California, Los Angeles.

Li, L-C., & O'Neil, H. F., Jr. (1995). *The effects of effort and worry on distance learning*. Manuscript submitted for publication. National Open University of Taiwan (LCL) and University of Southern California (HON).

O'Neil, H. F., Jr. (1972). Effects of stress on state anxiety and performance in computer-assisted learning. *Journal of Educational Psychology, 63,* 472-481.

O'Neil, H. F., Jr., Baker, E. L., Jacoby, A., Ni, Y., & Wittrock, M. (1990). *Human benchmarking studies of expert systems* (Report to DARPA, Contract No. N00014-86-K-0395). Los Angeles: University of California, Center for the Study of Evaluation/Center for Technology Assessment.

O'Neil, H. F., Jr., Baker, E. L., & Matsuura, S. (1992). Reliability and validity of Japanese trait and state worry and emotionality scales. *Anxiety, Stress, and Coping, 5,* 225-239.

O'Neil, H. F., Baker, E. L., Ni, Y., Jacoby, A., & Swigger, K. M. (1994). Human benchmarking for the evaluation of expert systems. In H. F. O'Neil, Jr., & E. L. Baker (Eds.), *Technology assessment in software applications* (pp. 13-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

O'Neil, H. F., Jr., & Brown, R. S. (1995). Item format and self-regulation. *Investigating the link between form and process in performance assessment.* Manuscript in preparation, University of Southern California/CRESST (HON), University of California, Los Angeles/CRESST (RSB).

O'Neil, H. F., Jr., & Richardson, F. C. (1977). Anxiety and learning in computer-based learning environments: An overview. In J. Seiber, H. F. O'Neil, Jr., & S. Tobias (Eds.), *Anxiety, learning and instruction (*pp. 133-146). New York: Lea/Wiley.

O'Neil, H. F., Jr., & Sugrue, B. (in press). Effects of motivational interventions on NAEP mathematics performance. *Educational Assessment*.

O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1992). *Final report of experimental studies on motivation and NAEP test performance (Report to NCES*, Grant No. RS90159001). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Paris, S. G., Cross, D. R., & Lipson, M. Y. (1984). Informed strategies for learning: A program to improve children's reading awareness and comprehension. *Journal of Educational Psychology, 76,* 1239-1252.

Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. In M. Maehr & C. Aimes (Eds.), *Advances in motivation and achievement: Motivation enhancing environments* (Vol. 6, pp. 117-160). Greenwich, CT: JAI Press, Inc.

Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82,* 33-40.

Pintrich, P. R., & Garcia, T. (1991). Student goal orientation and self-regulation in the college classroom. *Advances in motivation and achievement* (Vol. 7, pp. 371-402). Greenwich, CT: JAI Press, Inc.

Pressley, M., & Afflerbach, P. (1995). What readers can do when they read: A summary of the results from the on-line self-report studies of reading. In M. Pressley & P. Afflerbach, *Verbal protocols of reading: The nature of constructively responsive reading* (pp. 31-82). Hillsdale, NJ: Lawrence Erlbaum Associates.

Royer, J. M., Cisero, C. A., & Carlo, M. S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63*, 201-243.

Spielberger, C. D. (1972). Anxiety as an emotional state. In C. D. Spielberger (Ed.), *Anxiety: Current trends in theory and research* (Vol. 1, pp. 23-49). New York: Academic Press.

Spielberger, C. D. (1975). Anxiety: State-trait process. In C. D. Spielberger & I. G. Sarason (Eds.), *Stress and anxiety* (Vol. 1, pp. 115-143). Washington, DC: Hemisphere.

Spielberger, C. D. (1983). *Manual for the State-Trait Anxiety Inventory (STAI) (Form Y)*. Palo Alto, CA: Consulting Psychologists Press.

Tobias, S., & Everson, H. (1995, April). Development and validation of an objective measure of metacognition. In W. E. Montague (Chair), *Issues in metacognitive research and assessment*. Symposium conducted at the annual meeting of the American Educational Research Association, San Francisco.

Turing, A. M. (1988). Computing machinery and intelligence. In A. Collins & E. E. Smith (Eds.), *Readings in cognitive science* (pp. 6-19). San Mateo, CA: Morgan Kaufmann.

Weinstein, C. F., Palmer, D. R., & Schultz, A. C. (1987). *LASSI. The Learning and Study Strategies Inventory.* Clearwater, FL: H & H Publishing Company.

Yap, E. G. (1993). *A structural model of self-regulated learning in math achievement.* Unpublished doctoral dissertation, Los Angeles, University of Southern California.

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81,* 329-339.

Zimmerman, B. J., & Martinez-Pons, M. M. (1986). Development of a structured interview for assessing use of self-regulated learning strategies. *American Educational Research Journal, 23,* 614-628.

Zimmerman, B. J., & Martinez-Pons, M. M. (1988). Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology, 80,* 284-290.

Zimmerman, B. J., & Martinez-Pons, M. M. (1990). Student differences in self-regulated learning: relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology, 82,* 51-59.

Appendix
Self-Assessment Questionnaire

Directions. A number of statements which people have used to describe themselves are given below. Read each statement and indicate how you thought during the test. Find the word or phrase which best describes how you thought and circle the number for your answer. There are no right or wrong answers. Do not spend too much time on any one statement. Remember, give the answer which seems to describe how you thought during the test.

		Not at all	Some-what	Moder-ately so	Very much so
1.	I was aware of my own thinking.	1	2	3	4
2.	I checked my work while I was doing it.	1	2	3	4
3.	I attempted to discover the main ideas in the test questions.	1	2	3	4
4.	I tried to understand the goals of the test questions before I attempted to answer.	1	2	3	4
5.	I was aware of which thinking technique or strategy to use and when to use it.	1	2	3	4
6.	I corrected my errors.	1	2	3	4
7.	I asked myself how the test questions related to what I already knew.	1	2	3	4
8.	I tried to determine what the test required.	1	2	3	4
9.	I was aware of the need to plan my course of action.	1	2	3	4
10.	I almost always knew how much of the test I had left to complete.	1	2	3	4
11.	I thought through the meaning of the test questions before I began to answer them.	1	2	3	4
12.	I made sure I understood just what had to be done and how to do it.	1	2	3	4
13.	I was aware of my ongoing thinking processes.	1	2	3	4
14.	I kept track of my progress and, if necessary, I changed my techniques or strategies.	1	2	3	4

15. I used multiple thinking techniques or strategies to solve the test questions.	1	2	3	4
16. I determined how to solve the test questions.	1	2	3	4
17. I was aware of my trying to understand the test questions before I attempted to solve them.	1	2	3	4
18. I checked my accuracy as I progressed through the test.	1	2	3	4
19. I selected and organized relevant information to solve the test questions.	1	2	3	4
20. I tried to understand the test questions before I attempted to solve them.	1	2	3	4

Scoring Key

Scales	Items
Awareness	1, 5, 9, 13, 17
Cognitive Strategy	3, 7, 11, 15, 19
Planning	4, 8, 12, 16, 20
Self-Checking	2, 6, 10, 14, 18

AWARENESS

1. I was aware of my own thinking.
5. I was aware of which thinking technique or strategy to use and when to use it.
9. I was aware of the need to plan my course of action.
13. I was aware of my ongoing thinking processes.
17. I was aware of my trying to understand the test questions before I attempted to solve them.

COGNITIVE STRATEGY

3. I attempted to discover the main ideas in the test questions.
7. I asked myself how the test questions related to what I already knew.
11. I thought through the meaning of the test questions before I began to answer them.
15. I used multiple thinking techniques or strategies to solve the test questions.
19. I selected and organized relevant information to solve the test questions.

PLANNING

4. I tried to understand the goals of the test questions before I attempted to answer.
8. I tried to determine what the test required.
12. I made sure I understood just what had to be done and how to do it
16. I determined how to solve the test questions.
20. I tried to understand the test questions before I attempted to solve them.

SELF-CHECKING

2. I checked my work while I was doing it.
6. I corrected my errors.
10. I almost always knew how much of the test I had left to complete.
14. I kept track of my progress and, if necessary, I changed my techniques or strategies.
18. I checked my accuracy as I progressed through the test.