Differential Effects of Question Formats in Math Assessment on Metacognition and Affect

CSE Technical Report 449

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

Richard S. Brown
CRESST/University of California, Los Angeles

December 1997

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-6511
(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.

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

Acknowledgments

The authors wish to thank the following individuals for their participation in this study: Drs. Dale Carlson and Sue Bennett, California State Department of Education; Ms. Linda Murai and Ms. Darby Williams of Sacramento County Office of Education; Dr. Michael Seltzer, Ms. Katharine Fry, and Ms. Josie Bain of UCLA/CRESST; Dr. Winnie Young of CTB/McGraw Hill; and the test administrators and participants in the study.

DIFFERENTIAL EFFECTS OF QUESTION FORMATS IN MATH ASSESSMENT ON METACOGNITION AND AFFECT

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

Richard S. Brown
CRESST/University of California, Los Angeles

Abstract

This study investigated the effect of question format on metacognitive and affective processes of children in the context of a large-scale mathematics assessment program. Mathematical items were presented in both multiple-choice and open-ended question formats to eighth-grade students (N = 1,032) as part of the California Learning Assessment System (CLAS). Metacognition and affect were measured following each format for males and females of various ethnic groups. Results indicate that open-ended and multiple-choice formats have differential effects. Open-ended questions induced more cognitive strategy usage, less self-checking, and greater worry than did multiple-choice questions. These effects did not vary substantially as a function of gender and ethnicity.

In an effort to improve upon traditional measures of student performance, educational researchers have investigated alternate ways of measuring student achievement. Referring to these measures as performance assessments, alternative assessments, authentic assessments, and several other names, researchers hope to better capture student knowledge by allowing students to express their thinking and problem-solving strategies—to show what they know, with less emphasis on selecting a correct response from a set of alternatives (Baker, O'Neil, & Linn, 1993; Baxter & Shavelson, 1994; Herman, Aschbacher, & Winters, 1992).

Some argue that the performance-oriented question format also challenges students to think critically and allows students opportunities to draw upon prior knowledge and relevant skills to solve problems (Herman, Klein, Heath, & Wakai, 1994; Miller & Legg, 1993). Another claim is that these assessments stimulate

students to engage in complex thinking and thus reflect higher standards of excellence than traditional multiple-choice testing formats. However, the introduction of such assessment approaches raises questions of possible disparate impact on various ethnic groups (Baker & O'Neil, 1994; Winfield & Woodard, 1994).

With respect to equity, it is a guiding principle (Baker & O'Neil, 1995) that assessment should promote open access to educational services for all students without regard to their socioeconomic class, religion, gender, ethnicity, or primary language. If tests (or other assessments) motivate disadvantaged students less than others, and the importance and frequency of such assessments increase, then gaps in performance will also increase. Moreover, for all non-standard-English speakers, the dependence of assessment performance tasks on explanation, writing skill, and extended communication may create added difficulty (Baker & O'Neil, 1994). Finally, these assessments may interact with differences in students' instructional or opportunity-to-learn experiences and with the ethnic backgrounds of learners.

For example, one important equity issue is the extent to which the smaller number of problems or tasks used in performance assessments (Gao, Shavelson, & Baxter, 1994) can serve the broad diversity of students. Due to time constraints, most performance assessments consist of only a few tasks. In a generalizability context, without such domain specifications such few tasks lead to low reliability (Gao et al., 1994). An additional "cost" of a limited set of assessment tasks is the high likelihood that the assessment might include content to which some children may have had low exposure or interest.

Writers in the areas of ethnicity and performance (Miller-Jones, 1989; Steinberg, Dornbusch, & Brown, 1992) and ethnicity and motivation (Baker & O'Neil, 1995; Ogbu, 1978, 1992) have cited differences in student motivational characteristics which may affect their performance. For the most part, these assertions have been speculative, since limited performance assessment data are available as yet on ethnic or gender differences.

With respect to gender, previous research with high school and college-age students has found that females perform better than males on some performance-based assessments (e.g., essays) in comparison to multiple-choice formats (Bolger & Kellaghan, 1990; Breland, Danos, Kahn, Kubota, & Sudlow, 1991; Breland & Griswold, 1981, 1982; Bridgeman, 1989; Mazzeo, Schmitt, & Bleistein, 1992; Peterson & Livingston, 1982).

Explanations for these gender differences vary from genetic to social causes (see Mazzeo et al., 1992). Though some researchers have argued that males and females possess different abilities with regard to problem solving (e.g., greater verbal fluency for females, Breland et al., 1991), it has also been shown that the varying formats in alternative assessments induce different approaches to problem solving (Herman et al., 1994). The California Learning Assessment System[1] (CLAS) (California State Department of Education, 1993b), a statewide effort designed to assess student competencies in a variety of content areas and make comparisons to California's statewide content standards, used two item formats—performance oriented (or open ended) and selection (or multiple choice)—in assessing mathematical competency among California students.

Herman et al. (1994) reported that students employed different lines of reasoning in dealing with the two formats. Generally, students were more likely to use a trial-and-error or guessing approach for multiple-choice items but to use a mathematical line of reasoning for open-ended items. In addition, students perceived different criteria for successful performance on multiple-choice test items than they did for open-ended items. They noted that open-ended items emphasized the use of explanatory materials such as graphs and charts rather than focusing on algorithms and correct answers. Further, students held different expectations for multiple-choice versus open-ended test items. For example, more students mentioned the importance of the quality and depth of their response in the open-ended condition than in the multiple-choice condition. Understanding these different perceptions and the cognitive approaches students utilize to solve items of different formats may shed light on the nature of performance differences hypothesized between males and females and among various ethnic groups.

We have defined metacognition as the process by which individuals think about their own thinking in order to develop strategies to solve problems. This process has been broken down further into the subcategories of planning, self-checking, awareness, and cognitive strategy (O'Neil & Abedi, 1996; O'Neil, Sugrue, Abedi, Baker, & Golan, 1992; O'Neil, Sugrue, & Baker, 1996). (For complementary views of metacognition, see Borkowski & Muthukrishna, 1992; Borkowski & Thorpe, 1994; Everson, Smodlaka, & Tobias, 1994; Garcia & Pintrich, 1994; Paris, Cross, & Lipson, 1984; Pintrich & DeGroot, 1990; Zimmerman, 1994.) Further, we view

[1] The California Learning Assessment System has since been terminated and the state of California is revising its statewide assessment system.

metacognition, effort, and worry from a state-trait perspective (Spielberger, 1975). For example, state metacognition is defined as a transitory state in intellectual situations that varies in intensity, changes over time, and is characterized by planning, self-monitoring, changing cognitive/affective strategies, and self-awareness. This study will investigate the impact of item format on state metacognition.

Previous research suggests that cognitive strategies and self-checking behaviors are part of a series of state metacognitive learning behaviors that can enhance learning (Yap, 1993). Further, these strategies are all goal oriented, are intentionally invoked, require student effort, and are specific to the testing or learning situation (Weinstein & Meyer, 1991). Given the situational specificity of cognitive strategies usage, we were interested in how the specific issue of test-item format influences cognitive strategy utilization.

Finally, because metacognitive behavior is an effortful student activity in a testing situation, we sought also to determine the impact of different item formats on students' effort and worry in the context of CLAS math achievement.

Hypotheses

Building upon previous research, we anticipated several relationships between test-item format, gender, ethnicity, and metacognitive activity among students. Each is explicitly set forth below.

Guided by prior research findings (O'Neil et al., 1992), we anticipated that student math performance would vary as a function of the student characteristics of gender and ethnicity. We expected males to perform better than females as the content is mathematics, and we expected Asian and White students to perform better than African American and Latino students.

In general, we expected levels of cognitive strategy use, self-checking, worry, and effort to be higher among females than males, and Asians and Whites to exhibit more metacognitive behavior than African Americans and Latinos (O'Neil et al., 1992).

In addition, we believed test-item format would influence students' state metacognitive activity. We hypothesized that different item formats induce different levels of metacognitive activity among students. We expected that the

open-ended format would engage students in more cognitive strategy use and self-checking activity than would the multiple-choice format.

Given students' unfamiliarity with open-ended performance assessments, we expected that the open-ended item format would require more student effort and engender more worry than the multiple-choice format.

Finally, because researchers have expressed concern about the disparate impact on ethnic minorities of novel testing approaches (Winfield & Woodard, 1994) and about issues of inequitable opportunities to learn tested material among the disadvantaged groups (Linn, Baker, & Dunbar, 1991), we anticipated differential effects among student groups of varying ethnicity. We expected that the influence of item format on metacognition, effort, and worry would be more pronounced for African American and Latino students than for White and Asian students.

Methodology

Participants

The initial sample of subjects in this study was comprised of 1,480 eighth-grade students from 70 classes at 14 California middle or junior high schools. Some students were later excluded from our analysis. Of the initial group of 1,480 subjects, 56 students either failed to answer the ethnicity question, or answered with a category outside the four major ethnic categories of interest (i.e., White/Anglo; Black/African American; Hispanic/Latino; Asian or Pacific Islander). Another school administered the questionnaires to classrooms comprised of limited-English-proficient students (n = 18). These students were also excluded from the analyses.

In addition, one school administered the questionnaires in English classrooms rather than in math classrooms as prescribed in our instructions. We specified math classrooms over classrooms of other subjects because, in addition to collecting data at the student level, we were also interested in collecting data at the classroom level as it pertained to mathematics, so as to consider the multilevel nature of the data (see Burstein, 1980; Seltzer, 1994) in additional analyses not reported here. Consequently, these subjects (n = 107) were not included in this study. Additional subjects (n = 267) were dropped from the analyses due to incomplete data regarding which of the CLAS math assessment forms was used. As discussed below, evidence that perhaps some forms were easier than others required that form be used as a covariate in some of the analyses.

The final sample used in our analyses consists of 1,032 subjects from 12 schools in 59 classrooms. The sample is comprised of 527 males (51.1%) and 505 females (48.9%). The ethnic composition of this sample includes 440 White (42.7%), 88 African American (8.5%), 267 Hispanic/Latino (25.9%), and 237 Asian/Pacific Islander (22.9%). However, for the analyses concerning the open-ended math performance scores, the sample size was reduced to 335 since not all of the subjects in the sample were scored by the state on this measure. Subjects were randomly selected within each school for open-ended scoring in order to achieve a school-level CLAS score. Further, only one of the two opened-ended items was chosen for scoring randomly.

Procedure

As part of a planned research and development component for CLAS, a number of junior high schools were contacted by California State Department of Education personnel and asked to participate in the current study. Participation was voluntary for the school districts and schools, but not for the students as it was considered part of the statewide testing program.

Our questionnaires were administered within a few days to a few weeks following completion of the statewide CLAS math assessment. The CLAS math assessment administration took the entire math period. Thus, our measures of metacognition, effort, and worry were designed to be administered the next time the math class met. As the CLAS administration days varied from Monday through Friday, the next administration time varied from a minimum of 1 day (e.g., CLAS administered on a Monday and our measures administered on Tuesday) to a maximum 3-day period (CLAS administered on a Friday and our measure administered on the subsequent Monday). Due to a breakdown in communications, some classes experienced up to a 2-week delay between CLAS administration and administration of our measures. Unfortunately we did not ask for dates of administration on our measures, so a more precise accounting is not possible. Given the interest of the California State Department of Education (and thus the attention of the school principals) and the fact that we provided honoraria to teachers for assisting us, we think that our procedures with respect to time of administration were in general followed.

In the administration of our measures, eighth-grade students were instructed by a test administrator to rate themselves on state metacognition, effort, and worry

for two types of CLAS math questions: open-ended and multiple-choice items. The instructions focused on student reactions to the math problems of each type on the CLAS math assessment. A CLAS example of both formats was provided. In most cases, the questionnaires were administered in students' math classrooms by their usual math instructor who served as the CLAS administrator. The administrator read the following directions to the students:

> Today you will be participating in a statewide California study of students your age. To make sure that all students receive the same instructions, I will be reading them to you from a script.
>
> The intent of the present study is to collect information on your thoughts and feelings about the CLAS Middle Grades Performance Assessment. The state will use this information to improve the math assessment.
>
> We are particularly interested in your reactions to the open-ended problems as compared to the multiple-choice questions. Thus, in today's questionnaire we will ask you to respond to the same statements for each type of math question.
>
> As part of this study, you will answer questions about yourself and about mathematics. This will take about 40 minutes. You will not be allowed to ask questions while you are completing the questionnaire. If you have a question, save it until the end of the class and I will answer questions then.
>
> By doing the best you can, you will be making an important contribution.

The questionnaire took about 40 minutes to complete. After students completed the questionnaires, the test administrator gathered the study materials and returned them to the research personnel. Each administrator was provided an honorarium of $75.00 which could be used for classroom materials as compensation for his or her participation in the study.

Design

Because we were interested in the differential effect test-question format may have on the learning processes of various ethnic groups as well as between males and females, this study employed a three-factor mixed model design (Kirk, 1982). The study included ethnicity and gender as between-subjects factors and each of the metacognitive measures (cognitive strategy and self-checking) and affective variables (effort and worry) as within-subjects repeated measures.

Measures

Performance measures. The 1993 CLAS math assessment for eighth-grade students consisted of two open-ended questions and seven multiple-choice problems presented in eight different matrix-sampled forms (California State Department of Education, 1993a).

According to a Report of the Select Committee (Cronbach, Bradburn, & Horvitz, 1994), the accuracy of CLAS measures in 1993 as they relate to individual student scores in general (e.g., language arts or science) was in need of improvement. However, our math sample provides very reasonable, estimated standard errors of 0.5 and 0.6 for open-ended and multiple-choice math achievement scores, respectively. For the statewide data for the measures used in this study (eighth-grade mathematics), the estimated standard errors were sufficiently small (SE = .40) (Cronbach et al., 1994). Cronbach et al. also state that "an examination of other reports, some of them unpublished, indicates that CLAS scoring in Mathematics is achieving an accuracy level like that in other projects" (p. 36).

The multiple-choice questions were judged right or wrong depending upon selection of the appropriate answer. Students were allowed 15 minutes to answer all seven multiple-choice questions. A sample multiple-choice question is:[2]

> When a number is divided by 7, the remainder is 4.
>
> What is the remainder when twice that number is divided by 7?
>
> ___ A. 1 ___ B. 2 ___ C. 3 ___ D. 4

By contrast, the math open-ended problems allowed the students to express their manner of thinking and the process by which they came up with their answers, rather than focusing solely on the final answer as the object of assessment. Students were allowed 15 minutes to answer each open-ended problem. A sample open-ended problem is:

> Last year **Eat It Up Burgers** employed 5 workers for 5 hours a day. They claimed they served 4 million burgers last year. Is this a reasonable claim? Explain your answer.

Thus, the CLAS math assessment assessed math knowledge not only in the traditional multiple-choice format but also in the more performance-based open-

[2] The correct answer is A.

ended assessment format. Scores on the open-ended questions ranged from 0 to 4. Although two open-ended questions were asked of each student, only one was randomly scored by CTB/McGraw-Hill due to financial reasons. A multiple-choice performance measure was created by summing the number of correct multiple-choice responses to the seven multiple-choice questions (with score totals ranging from 0 to 7).

All items were scored in accordance with CLAS standards (California State Department of Education, 1993b), and items and scores were provided to us by the California State Department of Education and CTB/McGraw-Hill. Mean values for the raw score multiple-choice questions and the open-ended problem are provided in Table 1.

Table 1

Mean Raw Score Math Performance

Variable	Mean	*SD*
Multiple-choice (N = 1,032)	3.64	1.78
Open-ended (n = 335)	1.81	.90

Statistical analyses were conducted to determine any differences among the eight forms of the assessment (see Table 2). With respect to the multiple-choice questions, the forms were found not equivalent. Thus, form was used as a covariate for all analyses involving the multiple-choice measure. On the open-ended questions, performance across all 16 items (two items on each of the eight forms) was quite low, but a single factor analysis of variance test indicated no difference in performance as a result of test booklet, $F(7,327) = 0.57$, $p > 0.78$. Therefore, the form variable was not used as a covariate in the analysis concerning open-ended math performance.

Table 2

Mean Performance Measure by Form

	Multiple-choice (N = 1,032)		Open-ended (n = 335)	
Form	Mean	*SD*	Mean	*SD*
1	3.21	1.65	1.79	.95
2	3.21	1.57	1.63	.87
3	3.94	1.72	1.86	1.00
4	4.22	1.72	1.85	.83
5	3.57	1.88	1.74	.89
6	3.40	1.89	2.00	.95
7	4.07	1.79	1.84	.83
8	3.57	1.74	1.79	.95

Metacognition and affective measures. The metacognitive processes of self-checking and cognitive strategy and the affective components of effort and worry were assessed via a state inventory (O'Neil & Abedi, 1996) comprised of four subscales (six items for each subscale), scaled from 1 to 4 (1 = Not at all; 4 = Very much so) for each of the two question formats. For the metacognitive subscales, sample questions include "I went over my answers" and "I reworded the assessment items so I could understand them better" for self-checking and cognitive strategy, respectively. The affective components of effort and worry also were assessed (O'Neil et al., 1992) under each testing condition. Sample items measuring these constructs include "I did not give up, even if the assessment was hard" for effort and "I was afraid I should have studied more for this assessment" for worry. Thus, in total, there are two sets of four scales: one each measuring cognitive strategy, self-checking, worry, and effort for each of the two question formats (open ended and multiple choice).

The instructions for these scales were as follows:

> A number of statements that people have used to describe themselves are given below. Read each statement and indicate how you thought or felt while doing the multiple-choice (open-ended) math questions on the California Learning Assessment System

mathematics assessment. Find the word or phrase that best describes your thoughts or feelings and circle 1, 2, 3, or 4 in your booklet. There are no right or wrong answers. Do not spend too much time on any one statement. Remember, give the answer that seems to describe how you thought or felt while doing the multiple-choice (open-ended) mathematics assessment questions.

The four scales were modified from the state inventory (O'Neil et al., 1996). The modifications resulted from an item sensitivity review by California State Department of Education personnel, and the addition of new some items and deletion of some old items. For example, for the cognitive strategy subscale, three items were added and two items were dropped. For the self-checking subscale, four items were added and six items were dropped. For both these scales, the new items reflected the assumed psychological processes of both open-ended and multiple-choice formats. The items dropped possessed the weakest psychometric properties of their respective subscales.

The worry and effort scales (O'Neil et al., 1992) were reviewed in the same manner. No changes were made in item content to the worry scale, but two items were dropped due to the sensitivity review and poor item characteristics. With respect to effort, two items were added and five items were dropped. As may be seen in Table 3, the general result of these changes was to improve the reliability of the scales.

Table 3

Reliabilities of Metacognition, Effort, and Worry Measures

	Number of items		Alpha		
	O'Neil et al. (1996)	Present study	O'Neil et al. (1996)	Present study open-ended	Present study multiple-choice
Cognitive strategy	5	6	.61	.64	.74
Self-checking	5	6	.64	.72	.77
Effort	5	6	.79	.78	.82
Worry	8	6	.76	.75	.83

The raw score means and standard deviations for the scales are provided in Table 4.

Table 4

Mean Raw Score Metacognition, Worry, and Effort (N = 1,032)

Variable	Mean	*SD*
Multiple-choice		
Cognitive strategy	15.69	3.66
Self-checking	16.17	3.84
Worry	12.59	4.44
Effort	19.64	3.71
Open-ended		
Cognitive strategy	16.21	3.35
Self-checking	16.04	3.65
Worry	13.21	4.39
Effort	19.77	3.31

In addition, all 24 items on these four scales were subjected to a confirmatory factor analysis to investigate the validity of each scale. This process was performed separately for responses under both the multiple-choice and open-ended conditions. In both conditions, only the six items purported to indicate each of the four latent constructs were allowed to load on that construct, and each of the latent constructs was set to be correlated. These analyses were performed using the LISCOMP structural equation modeling program (Muthén, 1987).

Under both item-format conditions, the resulting covariance matrix estimated from the proposed factor structure reproduced the sample covariance matrix rather well (Root Mean Square Error = 0.067 in the multiple-choice condition, $\chi^2(246) = 1185.30$, $p < .001$, and Root Mean Square Error = 0.061 in the open-ended condition). A rule of thumb is that root mean square errors of less than .10 are acceptable.

Scale Reliabilities and Intercorrelations

As was shown in Table 3, analyses of the eight 6-item scales indicate reasonable internal consistency reliability for most of the measures. For the multiple-choice question condition, Cronbach's alphas were .74, .77, .82, and .83 for

cognitive strategy, self-checking, worry, and effort, respectively. For the open-ended condition, Cronbach's alphas were .64, .72, .78, and .75 for the same scales, respectively. It is of interest that the reliability for each scale is lower in the open-ended condition than in its multiple-choice counterpart. However, in all but one case the reliability measures exceed .70. For the one scale that dropped below this level (open-ended, cognitive strategy), additional analyses revealed that no specific item on the scale is especially contributory to the lower overall internal consistency. Thus, the scale was retained in the subsequent analyses in its complete form.

The interrelationship between the scales and the performance measures is shown in Table 5. In general, the pattern of correlations is similar for both question-format conditions. The effect of worry is, as expected, negative for both forms and the effect of effort is positive but low for both forms.

Table 5

Reliabilities and Intercorrelations for Metacognition, Effort, and Worry with Math Performance

Scale	1	2	3	4	Multiple-choice performance (N = 1,032)	Open-ended performance (n = 335)
Open-ended (N = 1,032)						
1. Cognitive strategy	1.00				.07*	.04
2. Self-checking	.63**	1.00			.08*	.07
3. Worry	.14**	.10*	1.00		-.24***	-.20***
4. Effort	.51**	.54**	.01	1.00	.13***	.22***
Multiple-choice (N = 1,032)						
1. Cognitive strategy	1.00				.00	.02
2. Self-checking	.68**	1.00			.03	.04
3. Worry	.23**	.14**	1.00		-.28***	-.20***
4. Effort	.52**	.57**	.05	1.00	.11***	.16**

* $p < .05$. ** $p < .01$. *** $p < .001$.

Results

Math Performance Scores

In general, math performance scores were quite low. The overall mean for the multiple-choice measure was 3.64 (SD = 1.78), and for the open-ended measure the mean was 1.81 (SD = .90). These scores represent 52.0% and 45.25%, respectively, of the possible total.

To investigate gender and ethnic effects in the multiple-choice condition, we subjected the data to a 2 X 4 (gender by ethnicity) analysis of covariance. We used test form as a covariate to control for form effects on the multiple-choice performance measure. For the open-ended performance measure, a similar analysis was performed with minor exceptions. First, test form was not used as a covariate in the open-ended condition, as no form differences were previously found. Second, African American students were excluded from the open-ended performance analyses due to small numbers of African American students with scored values on the open-ended measures. These insufficient numbers were the result of the reduced total sample size for open-ended measures (n = 335), due to a decision by the state to randomly score only one of the two open-ended items and to score only a subset of students in each school on this measure. Thus, the analyses for this outcome utilized a 2 X 3 (gender by ethnicity) analysis of variance approach.

We found no significant gender differences in math performance on the multiple-choice items or the open-ended problems. There were, however, ethnic differences in both question formats, F (1, 1023) = 42.90, p < .001; F (2, 327) = 5.93, p < .01, for the multiple-choice and open-ended measures, respectively. Post-hoc multiple group comparisons (Scheffe procedure) revealed that in the multiple-choice condition, Asian (M = 4.10, SD = 1.67) and White (M = 4.05, SD = 1.65) students performed significantly better than Latino (M = 2.88, SD = 1.78) and African American students (M = 2.70, SD = 1.56). In the open-ended condition, Asian students (M = 2.10, SD = .99) scored significantly better than Latino students (M = 1.60, SD = .83).

Metacognition, Worry, and Effort

To determine the influence that item format had on these variables, and whether item format impacts the relationships of gender and ethnicity with cognitive strategy, self-checking, worry, and effort, each of the measures was

subjected to repeated measures analysis of variance with ethnicity and gender as between-subjects factors and item format (open-ended vs. multiple-choice) as a within-subjects factor. A specific interest here is the presence or absence of significant interactions between item format and gender and item format and ethnicity. There were no forms differences on these variables; thus, covariance was not used. For the sake of brevity, only significant findings will be discussed.

Cognitive strategy. For cognitive strategy, there was considerable support for our hypotheses. As expected, there was a significant main effect for gender, F (1, 1024) = 14.95, $p < .001$, with females indicating more use of cognitive strategy (M = 2.73, SD = 0.58) than did males (M = 2.59, SD = 0.58). In addition, there was a significant main effect for question format, F (1, 1024) = 23.52, $p < .001$, with the open-ended questions (M = 2.70, SD = 0.56) inducing more use of cognitive strategy than did multiple-choice questions (M = 2.61, SD = 0.61). Further, there was a significant gender-by-question format interaction, F (1, 1024) = 6.80, $p < .01$. Inspection of cell means indicated that, although females exhibited more cognitive strategy usage than males in both conditions and both groups indicated more in the open-ended condition than in the multiple-choice condition, the difference between males and females was larger in the open-ended condition.

Self-checking. For self-checking, there was also a significant main effect for gender, F (1, 1024) = 17.23, $p < .001$, with females indicating more self-checking (M = 2.76, SD = 0.63) than did males (M = 2.61, SD = 0.62). And again, there was a significant effect for question format, F (1, 1024) = 6.37, $p < .05$. However, for this measure, multiple-choice questions (M = 2.69, SD = .64) yielded greater self-checking than did open-ended problems (M = 2.67; SD = .61). This may indicate that different self-checking behaviors are more functional, or at least more utilized, under varying testing formats. This may also be a result of the novelty of the open-ended question type. Students may not know how to check their performance and answers in this unfamiliar testing format. In addition to gender and question-format effects, there was a significant main effect for ethnicity found for self-checking, F (3, 1024) = 3.48, $p < .05$. Post-hoc comparisons for this measure indicate that in the open-ended condition, Latinos (M = 2.57, SD = 0.63) showed significantly less self-checking than did Whites (M = 2.71, SD = 0.59) and Asian Americans (M = 2.74, SD = 0.60). There were no significant differences among the ethnic groups in the multiple-choice condition, however.

Worry. Results for worry likewise indicated gender and question-format effects. For worry, there was a significant main effect for gender, F (1, 1024) = 5.42, $p < .05$, and for question format, F (1, 1024) = 31.68, $p < .001$. Consistent with the literature (Hembree, 1988, 1990), females indicated more worry (M = 2.20, SD = 0.75) than did males (M = 2.10, SD = 0.72), and open-ended problems induced greater amounts of worry (M = 2.20, SD = 0.73) than did multiple-choice questions (M = 2.10, SD = 0.74). Additionally, a significant main effect for ethnicity was found, F (3, 1024) = 16.85, $p < .001$. Post-hoc comparisons revealed that Whites exhibited significantly less worry than did all the other ethnic groups in both the open-ended and multiple-choice conditions. No other comparisons were significantly different in either testing format.

Effort. Regarding effort, significant differences were found as expected between males and females, F (1, 1024) = 22.27, $p < .001$, and among the ethnic groups, F (3, 1024) = 5.23, $p < .01$. The main effect of gender indicates that females (M = 3.37, SD = 0.55) exhibited more effort than did males (M = 3.20, SD = 0.60). The cell means for the four ethnic groups show that African Americans indicated they asserted the least effort among the groups (M = 3.13, SD = 0.62), followed by Latinos (M = 3.22, SD = 0.58), Asian Americans (M = 3.33, SD = 0.56) and Whites (M = 3.33, SD = 0.59).

The interaction between item format and ethnicity, F (3, 1024) = 2.34, p = .07, though not significant at the .05 level, may be regarded as cause for concern regarding the potential disparate impact that alternative assessment procedures may have on some ethnic minorities. Post-hoc comparisons reveal that no ethnic group comparisons were statistically significant in the multiple-choice condition but there were significant differences in the open-ended testing format. African Americans (M = 3.09, SD = .62) indicated significantly less effort than Whites (M = 3.33, SD = .56) and Asians (M = 3.35, SD = .50). Due to the small number of African American students whose open-ended performance was scored, there were not equivalent math performance data to compare to the effort data.

Relationship of Metacognition, Effort, Worry, and Performance

The correlations between the metacognition, effort, and worry measures and the performance measures (shown in Table 5) show that, for both the open-ended and multiple-choice conditions, worry and effort are significantly related to performance. However, these relationships are weak. In the multiple-choice condition, less worry

leads to better performance ($r = -.28$, $p < .001$) and more effort leads to better performance ($r = .11$, $p < .001$). The same pattern holds for the open-ended condition ($r = -.20$, $p < .001$; $r = .22$, $p < .001$; for worry and effort, respectively). Neither cognitive strategy nor self-checking was significantly correlated with performance in either the multiple-choice or open-ended testing condition. Given the tendency of the open-ended question format to generate more worry than the multiple-choice format, we may see reduced performance as a function of increased worry among test takers in the open-ended condition. Likewise, for effort, if the open-ended condition reduces the effort some African American students put forth on the assessment, performance deficits may result.

Summary and Discussion

In summary, we found mixed support for our expectations with respect to gender and ethnic differences and math performance. There were no gender differences in performance in either the multiple-choice or open-ended testing condition, but there were ethnic differences. Asian and White students performed better than Latino and African American students in the multiple-choice condition, and Asians did better than Latinos on the open-ended test items. (The reader may recall that there were too few African American students to conduct this analysis for open-ended items with an African American sample. White students were not statistically different from either Asian or Latino students; White students' performance was midway between these two groups.)

Consistent with our expectations, the metacognition, effort, and worry measures showed gender differences, with females consistently reporting more cognitive strategy, self-checking, effort, and worry than did males. An explanation of these gender effects is twofold: Either females are more open to expressing their thoughts and feelings than males in the eighth grade, or females are better students in the eighth grade in general and thus exhibit greater levels of metacognition and effort. There is evidence for both explanations. In addition, ethnic differences were found for self-checking, effort, and worry. Also consistent with our expectations, for both of the state metacognitive processes (cognitive strategy and self-checking) and for worry, question format produced significant differences. Only effort did not show a format difference, although the format and ethnicity interaction approached statistical significance ($p = .07$).

Our findings do not indicate simply an elevation on all measures for the less familiar open-ended question type. The open-ended format induced more cognitive strategy and worry than did the multiple-choice format, but the self-checking measure reflected less of this behavior in the open-ended condition than in the multiple-choice condition.

Although in general open-ended questions generated more cognitive strategy behavior than did multiple-choice items, this effect was more pronounced for females than for males. No similar finding was found for the measures of self-checking, worry, or effort. In addition, contrary to our expectations, there were no differential item-format effects on any of the measures for different ethnic groups (i.e., no interactions with ethnicity were significant), although the effort measure showed some indication of a possible trend ($p < .07$) toward the open-ended format to induce less effort from African Americans than from the other groups vis-à-vis the multiple-choice items.

These results indicate that open-ended and multiple-choice question formats have differential effects on metacognition, effort, and worry processes in student math achievement, and, further, that the open-ended format, in general, leads to more cognitive strategy activity and increases worry. However, the results of this study do not show that these formats are differentially penalizing members of gender or ethnic subgroups. None of the interaction effects among ethnicity and format were significant, indicating that the differential influence of item format does not operate differently among members of ethnic subgroups. Though open-ended items induce more worry and cognitive strategy use, they do so for Whites, Asians, African Americans, and Latinos alike. A note of caution, however, may be in order. The interaction between item format and ethnicity with regard to the effort variable did suggest the possibility that open-ended items may induce less effort from African-American children relative to the White and Asian counterparts. More research with a larger sample of various ethnic minorities is encouraged to investigate further whether alternative assessment formats treat all students equitably and fairly.

These findings must be viewed within the limitations of the current study. First, the performance measures are quite limited in their length; only seven multiple-choice items and one open-ended item were scored. The observed range of scores on these performance measures was very restricted, thus impacting observable correlations of metacognition, effort, and worry with math performance.

Second, the sample is limited to eighth-grade students and is quite small in dealing with the open-ended math performance measures (n = 335). This could limit the generalizability of findings and result in the study failing to have the power to detect all but large effects within and among the various ethnic groups. The study also does not consider the disparate opportunities to learn and demonstrate math knowledge in the open-ended format among the various ethnic groups. Third, the effects, though significant, are relatively small in magnitude. Thus, their practical significance should be thoughtfully considered. Many of the differences are of slight to moderate magnitude, with effect sizes frequently ranging from .25 to .50 standard deviation units. However, we would expect that if the state metacognition measures were given simultaneously with the math performance measures, the effects would increase in magnitude as well as significance.

Nevertheless, the importance of this research is clear. Alternative approaches to assessment may well induce different cognitive and affective processing by students. But these effects may also operate differentially across gender and ethnic groups. With continuing advances and recent implementations of various performance assessment techniques, it is imperative for educational researchers to investigate the influence the form of the assessment has on the experience of all students during the assessment process.

References

Baker, E. L., & O'Neil, H. F., Jr. (1994). Performance assessment and equity: A view from the USA. *Assessment in Education, 1*(1), 11-26.

Baker, E. L., & O'Neil, H. F., Jr. (1995). Diversity, assessment, and equity in educational reform. In M. Nettles & A. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 69-87). Boston: Kluwer Academic Publishers.

Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*, 1210-1218.

Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research, 21*, 279-297.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27*(2), 165-174.

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.

Borkowski, J. G., & Thorpe, P. K. (1994). Self-regulation and motivation: A life-span perspective on underachievement. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulation of learning and performance* (pp. 45-73). Hillsdale, NJ: Lawrence Erlbaum Associates.

Breland, H., Danos, D., Kahn, H., Kubota, M., & Sudlow, M. (1991). *A study of gender and performance on Advanced Placement history examinations* (College Board Report No. 91-4; ETS RR No. 91-61). New York: College Entrance Examination Board.

Breland, H., & Griswold, P. (1981). *Group comparisons for basic skills measures* (College Board Report No. 81-6; ETS RR No. 81-21). New York: College Entrance Examination Board.

Breland, H., & Griswold, P. (1982). Use of a performance test as a criterion in a differential validity study. *Journal of Educational Psychology, 74*, 713-721.

Bridgeman, B. (1989). *Comparative validity of multiple-choice and free-response items on the Advanced Placement examination in biology* (College Board Report No. 89-2; ETS RR No. 89-1). New York: College Entrance Examination Board.

Burstein, L. (1980). Analysis of multilevel data in educational research and evaluation. In D. Berliner (Ed.), *Review of research in education* (Vol. 8, pp. 158-233). Washington, DC: American Educational Research Association.

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

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

Cronbach, L. J., Bradburn, N. M, & Horvitz, D. G. (1994). *Sampling and statistical procedures used in the California Learning Assessment System*. Report of the Select Committee, July 25.

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.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7*, 323-342.

Garcia, T., & Pintrich, P. R. (1994). Regulating motivation and cognition in the classroom: The role of self-schemas and self-regulatory strategies. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulation of learning and performance* (pp. 127-153). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47-77.

Hembree, R. (1990). The nature, effects and relief of mathematics anxiety. *Journal for Research in Mathematics Education, 21*(1), 33-46.

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

Herman, J. L., Klein, D. C., Heath, T. M., & Wakai, S. T. (1994). *A first look: Are claims for alternative assessment holding up?* (CSE Tech. Rep. No. 391). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Kirk, R. (1982). *Experimental design.* Belmont, CA: Wadsworth.

Linn, R., Baker, E. L., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Mazzeo, J., Schmitt, A., & Bleistein, C. (1992). *Sex-related differences on constructed-response and multiple-choice sections of advanced placement examinations: Three exploratory studies.* Princeton, NJ: Educational Testing Service.

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice, 12*(2), 9-15.

Miller-Jones, D. (1989). Culture and testing. *American Psychologist, 44*, 360-366.

Muthén, B. O. (1987). *LISCOMP. Analysis of linear structural equations with a comprehensive measurement model.* Mooresville, IN: Scientific Software.

Ogbu, J. (1978). *Minority education and caste.* San Diego, CA: Academic Press.

Ogbu, J. (1992). Understanding cultural diversity and learning. *Educational Researcher, 21*(8), 5-14.

O'Neil, H. F., Jr., & Abedi, J. (1996). Reliability and validity of a state metacognitive inventory: Potential for alternative assessment. *Journal of Educational Research, 89,* 234-245.

O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1992). *Report of experimental studies on motivation and NAEP test performance* (Final Report to the National Center for Education Statistics, Contract No. RS 90159001). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

O'Neil, H. F., Jr., Sugrue, B., & Baker, E. L. (1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3*, 135-157.

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.

Peterson, N., & Livingston, S. (1982). *English composition tests with essay: A descriptive study of the relationship between essay and objective scores by ethnic group and sex* (ETS Rep. No. SR-82-96). Princeton, NJ: Educational Testing Service.

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

Seltzer, M. H. (1994). Studying variation in program success: A multilevel modeling approach. *Evaluation Review, 18*, 342-361.

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.

Steinberg, L., Dornbusch, S. M., & Brown, B. B. (1992). Ethnic differences in adolescent achievement: An ecological perspective. *American Psychologist, 47*, 723-729.

Weinstein, C. E., & Meyer, D. K. (1991). Cognitive learning strategies and college teaching. *College Teaching: From Theory to Practice, 34*, 15-26.

Winfield, L. E., & Woodard, M. D. (1994). Assessment, equity and diversity in reforming America's schools. *Educational Policy, 8*(1), 3-27.

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. (1994). Dimensions of academic self-regulation. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulation of learning and performance* (pp. 3-21). Hillsdale, NJ: Erlbaum.

Appendix

Scoring Key
State Post Thinking Questionnaire – Open-Ended

Scales	Items
Cognitive Strategy	3, 7, 9, 12, 15, 20
Self-Checking	2, 5, 10, 14, 18, 21
Worry	4, 8, 13, 17, 19, 23
Effort	1, 6, 11, 16, 22, 24

Student Questionnaire

Directions for Open-Ended Problems: A number of statements that people have used to describe themselves are given below. Read each statement and indicate how you thought or felt while doing the open-ended math problems on the California Learning Assessment System mathematics assessment. Find the word or phrase that best describes your thoughts or feelings and circle 1, 2, 3, or 4 in your booklet. There are no right or wrong answers. Do not spend too much time on any one statement. Remember, give the answer that seems to describe how you thought or felt while doing the open-ended mathematics assessment problems.

		Not at All	Somewhat	Moder-ately So	Very Much So
1.	I concentrated as hard as I could when taking the assessment.	1	2	3	4
2.	I checked my work while I was doing it.	1	2	3	4
3.	When solving a problem, I tried more than one way to do it.	1	2	3	4
4.	I thought my score was bad, so everyone, including myself, would be disappointed.	1	2	3	4
5.	I went over my answers.	1	2	3	4
6.	I worked hard on the assessment even though it did not count.	1	2	3	4
7.	I reworded the assessment problems so I could understand them better.	1	2	3	4
8.	I was afraid I should have studied more for this assessment.	1	2	3	4
9.	I selected and organized relevant information to solve the assessment problems.	1	2	3	4
10.	I judged the correctness of my work.	1	2	3	4
11.	I put forth my best effort.	1	2	3	4
12.	I thought through the meaning of the assessment problems before I began to answer them.	1	2	3	4

	Not at All	Somewhat	Moder-ately So	Very Much So
13. I felt regretful about my performance on the assessment.	1	2	3	4
14. I asked myself, how well was I doing, as I proceeded through the assessment.	1	2	3	4
15. On difficult problems, I spent more time trying to understand them.	1	2	3	4
16. I kept working, even on difficult assessment problems.	1	2	3	4
17. I wasn't happy with my performance.	1	2	3	4
18. I corrected my errors.	1	2	3	4
19. I was concerned about what would happen if I did poorly.	1	2	3	4
20. When solving a problem, I translated the problem into a different form.	1	2	3	4
21. As I did the assessment, I asked myself questions to stay on track.	1	2	3	4
22. I tried to do my best on the assessment.	1	2	3	4
23. I did not feel very confident about my performance on the assessment.	1	2	3	4
24. I did not give up, even if the assessment was hard.	1	2	3	4