ED 435 626	TM 030 224
AUTHOR	Suzuki, Kyoko; Harnisch, Delwyn L.
TITLE	An Investigation on the Generalizability of Performance-Based Assessment in Mathematics.
PUB DATE	1996-04-00
NOTE	42p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Algebra; Foreign Countries; *Generalizability Theory; *High School Students; High Schools; Holistic Evaluation; *Mathematics; Mathematics Tests; *Performance Based Assessment; Reliability; Scoring
IDENTIFIERS	Japan; United States

ABSTRACT

This study examined the generalizability and dependability of a performance-based assessment in algebra. Four forms of a five-item test were constructed using different subsets of eight items based on attributes from task analysis. Subjects included 142 "algebra 2" students from 2 high schools in the Midwestern United States and 148 11th graders from 1 school in Japan. Students' responses were scored using a holistic scoring rubric from 0 to 4. Analyses of generalizability and dependability revealed that the four forms achieved moderate levels of generalizability, although they varied by school. Analyses of data from subsets of the items on the forms suggested that acceptable levels of generalizability could be achieved by using items if items were well chosen. The choice of subsets of the items did not affect the validity of the content coverage of the test. One U.S. school and one Japanese school showed more similarity than between two U.S. schools. Four appendixes contain a list of mathematics attributes, sample test items, a scoring guide, and a scoring example. (Contains 16 tables, 1 figure, and 18 references.) (Author/SLD)

Running head: THE GENERALIZABILITY OF PERFORMANCE-BASED ASSESSMENT

An Investigation on the Generalizability of Performance-based Assessment in Mathematics

Kyoko Suzuki
Delwyn L. Harnisch
Department of Educational Psychology
University of Illinois at Urbana-Champaign
210 Education Bldg.
1310 South Sixth St.
Champaign, IL 61820
phone: (217) 333 - 2245
fax: (217) 244 - 7620
email: k-suzuki@uiuc.edu

Paper presented at the 1996 annual meeting of the American Educational Research Association, April, 8 - 12, New York.

Abstract

This study examines the generalizability and dependability of a performance-based assessment in algebra. Four forms of a five-item test were constructed using different subsets of 8 items based on attributes from task analysis. Subjects included 142 “algebra II” students from two high schools in the midwestern U.S., and 148 eleventh graders from one school in Japan. Students' responses were scored using a holistic scoring rubric from 0 to 4. Analyses of generalizability and dependability revealed that the four forms achieved moderate levels of generalizability, although they varied by school. Analyses of data from subsets of the items on the forms suggested that acceptable levels of generalizability could be achieved by using items if items were well chosen. The choice of subsets of the items did not affect the validity of the content coverage of the test. One American school and one Japanese school showed more similarity than between two American schools.

An Investigation on the Generalizability of Performance-based Assessment in Mathematics

Alternative assessment has moved to center stage as the focus of assessment has changed in the past decade. The current movement toward alternative assessment has brought the need for constructing a new test theory for measuring achievement levels in large scale testing situations: (a) How can we assess achievement and understanding? (b) how can we measure the student's thinking processes? and (c) how can we measure the cognitive growth in an achievement test?

Since multiple-choice tests derive their value as educational indicators, the indicators are often confused with instructional goals, which has led to an overemphasis on indicators as an educational goal. The lack of correspondence between indicators and goals provides the motivation for alternative assessment which directly measures complex performance including more open-ended problems, essays, or hands-on activities. Performance-based assessment can be useful for measuring students' proficiency in solving complex mathematical problems, reasoning and communicating mathematically (Lane, Stone, Ankenmann, & Liu, 1994). However, it is not enough to assume that alternative assessment for complex learning and processes is more valid than multiple-choice tests. Selected criteria need to be addressed for evaluating new assessments in a theoretical framework of validity (Harnisch, 1994; Linn, Baker & Dunbar, 1991).

Task structure is a criterion to ensure a valid assessment of students' proficiency. One purpose of achievement tests is to allow students to display their thinking at their best as a result of instructional programs in schools. Therefore, assessment instruments need to be developed to measure variety of strategies in solving problems, reasoning skills, and communications mathematically (Lane, 1993).

Intertask consistency is another criterion for validity evidence of performance-based assessments in evaluating the extent to which the results from an assessment lead to valid generalizations to a broadly-defined domain. Performance-based assessment tasks should be consistent not only across raters and similar tasks, but also across tasks that vary in content or format but that represent the same domain (Dunbar, Koretz, & Hoover, 1991). Thus, the unidimensionality of tasks in a test may be jeopardized. The intertask consistency in both science and writing performance assessments indicated that the generalizability of individual-level scores derived from assessment consisting of three to five tasks was questionable (Shaveleson, Baxter, & Pine, 1992; Miller & Crocker, 1990).

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 1983; Shavelson & Webb, 1991) investigates the degree to which results of performance assessment can be generalized to infer students' abilities. Some studies pointed out that the generalizability across tasks in performance-based assessment were limited (e.g. Shavelson, et al., 1992). However, the limited degree of generalizability across tasks needs to be taken into account in the design of an assessment program, generally, either by increasing the number of performance-based tasks or by using a matrix sampling design (Linn, et al., 1991).

In this study, the generalizability theory is applied to assess reliability and dependability of derived scores. Three questions are discussed:

(1) What is the minimum number of items in a test needed to achieve suitable generalizability?

(2) How can we select tasks to obtain sufficient levels of the generalizability of performance-based assessment for measuring students' achievement levels of mathematics learning?

(3) What criteria of the task quality are needed for performance-based items to attain sufficient generalizability?

Method

Participants

Data were collected from 290 students in three high schools during November, 1994 to January 1996. One hundred forty-two "Algebra II" students (10 - 12th graders) participated from two schools in the midwestern U.S.: one in a small city, and the other in a suburb. One hundred forty-eight 11th graders participated from one school in a suburb in Japan. These two U.S. schools were planned to be combined as an American sample for a comparison with a Japanese sample. However, the statistical analyses are reported individually in this study, because statistical analyses showed large differences between the two schools in the U. S..

The reason for including a sample from Japan was to compare a cultural effect that was experienced in performance-based tasks in mathematics. Performance-based tests are relatively new test formats for many American students. On the other hand, there is no use of multiple-choice tests in mathematics for college entrance examinations in Japan; therefore, Japanese high school students are trained how to write their answers mathematically in school programs.

Teachers usually refer to it as "mathematical technical writing" or "math composition." Mathematical writing is stressed in math lessons as a technique for entrance examinations. Math writing is also very difficult for many Japanese high school students, and both teachers and students spend time and much effort to acquire these skills. The problem with math composition in Japan is that the motivation is derived as a "technique" for entrance examinations, but not from educational goals. Because Japanese students seem to be more familiar with performance-based tasks, a sample was included to compare generalizability of items with samples from the U.S..

Materials

Performance-based tasks.

Eight tasks in algebra were chosen from similar content areas and were grouped into 4 forms having 5 items each based on the Tatsuoka's attribute chart in Appendix A (Tatsuoka, 1992). The reasons for using four different types of forms were (a) feasibility in a classroom hour (approximately 45 minutes), and (b) detecting which items mostly likely have higher generalizability. The time constraints must be considered for administering a performance-based, achievement test in a classroom. For minimizing the effect of speededness, five items were chosen for each form to examine which form and which items could achieve the highest generalizability. All items are shown in Appendix B. Item composition in each form is shown in Table 1. The mathematics ability measured in each item, the attributes' classification per item, and the attributes measured on each form are shown in Table 2, Table 3, and Table 4 respectively.

Five of 8 items were modified from publicly released SAT multiple-choice items, two items were developed as performance-based items, and one item was a typical textbook problem. As seen in Table 4, the forms were not strictly parallel; however, each form covered most of attributes equally. One reason for using five items as performance-based tasks modified from multiple-choice items was that most tasks on SAT were validated as reliable items. However, when these items were used as performance-based tasks, the constructs measured in each items may be different, because multiple-choice tasks measure only the presence or absence of the knowledge, whereas performance-based tasks measure different constructs such as mathematical knowledge, strategic skills and communication. Therefore, the five items modified from multiple-choice tasks were included to verify the task structure for performance-based tasks: Which items can be used as performance-based tasks as well, and which items are not appropriate?

Scoring rubric.

The scoring rubric in this study was adopted from the QUASAR[1] (Quantitative Understanding: Amplifying Student Achievement and Reasoning) project. The students' responses were scored from 0 to 4 scale using the holistic perspective considering three components; mathematical conceptual and procedural knowledge, strategic knowledge, and communication. The description of this rubric is shown in Appendix C.

[1] QUASAR (Quantitative Understanding: Amplifying Student Achievement and Reasoning) is a national project that seeks instructional programs in the middle-school grades that promote the acquisition of thinking and reasoning skills in mathematics (Silver, 1991). The project is directed at students attending schools in economically disadvantaged communities.

Procedure

The four forms of the tests were randomly assigned to students in one class period (approximately 45 minutes). Calculator use was allowed in the U.S., but not in Japan. The participants distributions per form per school were shown in Table 5.

Students' responses were scored by two trained raters using the scoring rubric described previously. The inter-rater reliability was about .9. The reason for this high inter- rater agreement was due to the working expriences between two raters. They were working together for years on the same project of performance-based assessment.

Scoring procedures in this study focused on assessing the reasoning and communicating skills for finding their answers rather than the final answers. The stress was on the process of finding answers and to communicating solution strategies with others in written verbal format. Therefore, a response could receive a "4" (the highest score) if the strategy and process were given sufficiently, even though the final answer was not correct. On the other hand, a response could be scored a "2" when a solution process was not given or poor, although the final answer was correct. Scoring examples are given in Appendix D.

Analysis of variance (ANOVA) was used to examine mean score differences, and Scheffe's post-hoc tests were used with a significance level of .05. First, school differences on total scores, form differences on total scores for all data, interaction effects between form and school were examined. Next, mean total scores at school level were compared for each item. For comparing mean item scores to examine item difficulty at each school level, repeated measure design ANOVA was used. Lastly, correlation analysis was conducted to examine correlations among items at each school level.

For the generalizability study, the computer program LIKERT (1994) was used to calculate point-biserial coefficients and Cronbach's coefficient alphas for each test form. Then, the computer program GENOVA (1982) was used to calculate generalizability and dependability coefficients for each form. To increase generalizability coefficients in each test form, some items were selected based on point-biserial coefficients. Then, generalizability and dependability coefficients were re-calculated by GENOVA for each revised form based on selected items.

Results

Comparison of Mean Scores: School, Form, Item.

School difference on mean total score was significant at the school level ($F(2, 287) = 46.06$, $p < .0001$, see Table 6). Therefore, the students' average achievement levels were different between the schools. Mean total scores on each form were not significantly different between the four forms ($F(3, 286) = 0.50$, $p = .6807$). There were no differences in difficulties by test form observed for all data from the three schools.

Because an interaction effect was observed between schools and forms ($F(6, 278) = 5.22$, $p < .0001$), difficulty levels of forms were examined at each school (see Table 7). Mean total scores on each form at School 1 were significantly different ($F(3, 43) = 5.23$, $p = .0036$). The mean total scores were C > A > B > D, which means Form C had the highest mean total scores and Form D had the lowest mean total scores. Form C and Form B, and Form C and Form D were significantly different in Scheffe's post-hoc test. At School 2, the difficulty of each forms was not significant ($F(3, 91) = 1.87$, $p = .14$, A > C > D > B). At School 3, the form difficulty was significant ($F(3, 144) = 3.65$, $p = .0142$, B > A > D > C). However, only Form B and Form C were significantly different. Based on this analysis, Form D might be slightly more difficult than other forms across schools. The effect of difficulty levels by test form was negligible.

Mean scores on each item at each school level were compared to examine difficulty levels among items. The difficulty among items was significantly different at each school level (F(7, 181) = 7.90, p < .0001, at School 1; F(7, 373) = 15.98, p < .0001, at School 2; F(7, 585) = 24.17, p < .0001, at School 3). The significantly different items at each school are shown in Table 8. It was notable that no cell of the Table 8 was shared by all schools, and that no cell was shared by two American schools. Based on this analysis, although item difficulty was slightly different among some items, there were no common tendency observed across schools. The item difficulty seemed dependent on schools and the instructional opportunity given to children.

Correlation Analysis

Correlation analysis was conducted to examine the relationship of mean scores among items. If two items were highly correlated, the two items measured the same or similar ability. Based on Table 9, Item 1 and 3, and Item 2 and 3 were significantly correlated at all schools. This result was consistent with the analysis of Attributes' chart by item in Table 3. However, the correlation among other items were dependent on schools. Item 4 did not have any significant correlation with other items, although it share some attributes with other items. This result showed that the students' performances were not consistent with the attributes in task analysis. All items were significantly correlated with the totals scores except Item 8 at School 1, which meant that all item scores were highly related to the total score, but Item 8 did not show this relation with the total score for School 1 students.

Generalizability Study and Decision Study

Computer program LIKERT (1994) was used to calculate point-biserial coefficients for each item along with coefficient alphas for each form by school. Alpha values for each form raged quite differently at each school level (see Table 10): 0 to .53 at School 1, .33 to .66 at School 2, and .30 to .78 at School 3. Form A tended to have lower alphas compared to others, which meant that Form A had relatively lower test consistency. Form D had lower alphas for American students (.22 at School 1 and .33 at School 2), meanwhile sufficient alpha level was attained for Japanese students (.78 at School 3). However, there were no tendencies observed in other forms between American students and Japanese students. In addition, which form attained higher test consistency depended on schools, although School 2 had relatively higher and more stable alphas across four forms.

Intertask consistency was examined using person by item (P x I) generalizability design for each form. This design examined differential student performance across items. The computer program GENOVA (1982) was used to estimate random-effect variances, generalizability coefficients, and dependability coefficients.

The random-effect variance estimates for each form by school are shown in Table 11. Because of zero alpha coefficient in Form A at School 1, both variance components of person and of item were not estimated. Across other forms, the variability due to person accounted for between 2% and 30% of the total variability: 2 % to 13 % at School 1, 6 % to 28 % at School 2, and 7 % to 30 % at School 3. School 1 relatively had lower values than other schools, which was consistent with the result that coefficient alphas for each form except Form C were lower in School 1 than in other schools.

The variability due to items ranged from 0 % to 36% of total variability: School 2 ranged from 0 % to 36 %, whereas School 1 ranged from 14 % to 26 %. The variance component for the P x I interaction represents the differential performance of students across tasks but it is also confounded with random error variance and variance due to systematic influences not included in the design. The variability due to the P x I interaction accounted for a large percentage of the total variability: between 56% and 84%.

For decision studies, a random effects design was used. Person by item (P x I) design, and item nested in person (I : P) design were used for calculating generalizability coefficients and dependability coefficients. The generalizability coefficients are for relative decisions in which the rank order of students is of interest. Therefore, any constant effect for all students are not considered as a factor of unreliability. Meanwhile, the dependability coefficients are for absolute decisions in which the absolute level of performance is of interest and fluctuations in mean scores of items are considered as a factor of unreliability. Generalizability coefficients and Dependability coefficients are reported in Table 12.

Suppose that a test contains nine tasks, the highest generalizability coefficients for each form range from .63 (Form A, School 2) to .86 (Form D, School 3) for (P x I) design. The highest dependability coefficients for each form range from .60 (Form A, School 2) to .79 (Form D, School 3) for both (P x I) design and (I : P) design. The generalizability and dependability coefficients are affected by the range of student performances. Because the variance due to person for Form D at School 3 was relatively large (30 % of the total variance), Form D at School 3 attained the higher generalizability and dependability coefficients. On the other hand, the person variance was small (14 % at School 2) in Form A, which contributed to the lower generalizability and dependability coefficients. This result was consistent with the result of Cronbach's coefficient Alpha. Form D attained a sufficient level of generalizability at School 3, whereas very low at both School 1 and School 2. Form B and Form C had similar levels of generalizability across schools. Note that the difference between the generalizability and dependability coefficients was small when the variance due to item was small.

This statistical analysis demonstrated that the possibility of attaining higher generalizabiltiy for each test form depended on school samples. There was no form verified as a less reliable test. Additional study is needed to determine which items contributed higher generalizability and dependability, or which items should be deleted to attain high intertask consistency in a test.

Generalizability Study for Selected Items' Form

The original four forms attained relatively low generalizability, although it varied by school levels. One objective of this study was to examine a method to attain higher generalizability in performance-based tests with a relatively small number of tasks. Therefore, a subset of items from each form was selected and examined their generalizability and dependability coefficients. Three items were selected in each form based on higher point-biserial coefficients, which are an indicator of item-test consistency.

The computer program LIKERT was used to calculate coefficient alphas for the selected items' forms (see Table 13). Compared to the alphas of original forms, these values were increased except Form D at School 3 (from .78 for original to .72 for selected). Form A tended to have lower alpha levels.

The computer program GENOVA was used to calculate variance estimates, and the generalizability and dependability coefficients for the selected items' forms. The random-effect variance estimates for each selected items' form by school are shown in Table 14. The variability due to person increased from original forms, between 12 % to 54 % of the total variability. Since the person variability affects the generalizability and dependability coefficients, the higher values of these coefficients were expected for selected items' forms. It was notable that selected items' Form A still showed that the large percentage of the total variance was the variance due to the (P x I) interaction, between 59 % and 84 %, which represented an error term.

For decision studies, a random effects design was used for selected items' forms. Person by item (P x I) design, and item nested in person (I : P) design were used for calculating generalizability coefficients and dependability coefficients. Generalizability coefficients and Dependability coefficients are reported in Table 15.

Assuming nine tasks in a test, the highest generalizability coefficients for each form ranged from .78 (Form A, School 2) to .92 (Form B, School 1) for (P x I) design. The highest dependability coefficients for each form ranged from .74 (Form A, School 2) to .91 (Form B, School 1 and School 2) for both (P x I) design and (I : P) design. These values were significantly increased from the values for the original forms. This statistical analysis suggested that the sufficient levels of generalizability for performance-based tests with a small number of tasks could be attained if items in a test were well chosen. In addition, the generalizability and dependability were affected by students' samples.

The last step in this study was to determine which items should be selected. Table 16 showed the pattern of item selection at each school based on point-biserial coefficients, which represented item-test consistency. Based on this analysis, all items were chosen at least twice. Only Item 4 was supported by one school: The rest of items were supported by at least two schools. Item 4 and Item 7 had lower supports than others. Item 6 was supported at all occasions: Item 3 was highly supported, too. Item 2 was supported by all schools in form D, but no support in form C. On the other hand, Item 5 was supported by all schools in Form C, but none in form D. Item 1, 2, 3, 5, 6 were supported by all three schools in at least one form, which suggested that they were good candidates as performance-based tasks. School 2 and School 3 had exactly same patterns in item selection in Form C and Form D. This was a surprising result, because an American school and a Japanese school showed more similar pattern in item selection than between two American schools!

Discussion

Achievement levels were significantly different between the three schools. Difficulty levels by test form were not significantly different, although they varied among schools. Therefore, the effect of difficulty levels by test form in this study is negligible. Item difficulties varied among the schools with none of the differences common to all three schools. It is notable that none of the item difficulty levels were common for the two American schools, although some of them are common between an American school and the Japanese school. These results imply that no item can be detected as a more difficult item than others and that no form can be detected as a more difficult test.

Correlation analysis demonstrated that some of the items were significantly correlated at all school levels, which was consistent with the task structure. However, some items (e.g. Item 6 and Item 7) which were supposed to have similar task structure with others (such as Item 1, 2, or 3) did not show significant correlation at some schools. This result implies that actual students' performances may be different from expected students' performances or knowledge levels.

Original forms showed relatively low coefficient alphas, especially in Form A, although the alphas varied at school levels. The low alpha levels contributed relatively low generalizability and dependability coefficients for original forms. When a subset including three items was selected in each form based on higher point-biserial coefficients, the selected items' forms attained higher alphas, which contributed to attain higher generalizability and dependability coefficients, although Form A still attained less generalizability. This result suggests that the performance-based tests with small numbers of tasks can attain high generalizability if items are well chosen.

Item selection pattern showed interesting features. First, Item 6 and Item 7 have same attributes as task analysis. However, Item 6 is highly supported, whereas Item 7 is less supported. Because the item selection is based on intertask consistency which is affected by students' performances, the result may be practical evidence of the difference between actual students' performances and expected performances. Although the task structure is similar for some items, students' performances may vary. To examine the differences of students' performance among tasks which have similar task structure, qualitative analyses of students' responses are necessary such as protocol analyses.

There observed no common traits across schools for item selection patterns. In fact, the two American schools did not show much commonality. Instead, an American school and the Japanese school showed the identical item selection patterns in Form C and Form D. What does this imply? Because the item selection is based on item-test consistency levels, the selection is affected by students' performance. This result relates the generalizability and dependability coefficients. These coefficients are also affected by levels of students' performances. Therefore, this result may be caused by similar students' achievement levels on these performance-based tasks between School 2 and School 3. Other associated features with this result may be the sample size and students' characteristics. School 1 had only 48 participants in total, which means 11 to 12 responses on each test form; therefore, the statistical power is not large enough to generalize some statistical traits. No responses to a task were often observed at School 1, which may contribute to low intertask consistency.

Speededness effects needs to be evaluated for performance-based tests. If the assessment is speeded, the validity of the score interpretation is questionable. The students' omission of items may be an indicator of speededness(Lane, Liu, Stone, & Ankenmann, 1993). Frequent omissions on Item 8 at School 1 might be an indicator of speededness. However, the observation of test administration in other three schools in February, 1996, revealed that the speededness effects was negligible in this study. In fact, 45 minutes were too much to complete five tasks for most students, although they did not answer all the questions in a test. Some students said that they did not write anything in a task because they did not know what to write for the question. Some students said that they simply "gave up to answer" because the task seemed too difficult for them. Therefore, the frequent omission of tasks in this study can be considered as an indicator of seriousness of examinees.

Item 1, 2, 4, 5, 6 were modified form SAT multiple-choice items. Based on the analyses of this study, four of five items can be used as performance-based items as well, whereas Item 4 was less supported. Multiple-choice tasks can be modified as performance-based tasks, and they can distinguish students' proficiency levels in various constructs as well. This implies that performance-based tasks can be developed from multiple-choice type context and found to be useful as well. However, which items are more preferable as multiple-choice type, and which items are appropriate as performance-based tasks? And why? These questions should be examined by qualitative analyses of students' responses as well as by statistical evidence.

Qualitative analyses of students' responses are necessary to examine what criteria of items are needed for attaining higher generalizability in a test. As mentioned previously, the difference between Item 6 and Item 7 is caused by students' performances, but not the task structure. In fact, Item 6 can be solved intuitively, such as "guess and check," whereas Item 7 requires more luck to find the answer by "guess and check" method. This tendency is also observed in other items, in which students can use "plug in numbers" method to find the answer. In fact, this method is effective when calculator is allowed in a test. The complexity levels of items should be examined qualitatively based on students' responses to determine the value as a performance-based task: Is it worth-while for students and raters to determine students' achievement levels? Students' performances on tasks can reveal level of learning: thinking strategies, reasoning processes, and communication skills. Qualitative analyses of students' responses raise some interesting issues for future research, (a) solution strategies, (b) complexity levels of tasks, and (c) calculator use in a test. These issues need to be examined in more detail (e.g. Suzuki & Harnisch, 1996).

Limitation of This Study

First, only tasks in algebra are used in these tests; therefore, the result can apply only for achievement tests in algebra. Second, the statistical analyses does not capture the lack of reliability due to the nature of the items. This issue can be examined by qualitative analyses of students' responses. Third, there is no evidence to examine the lack of reliability due to the difficulty of applying the scoring rubric. However, the rater reliability is grater than .9 in this study, which implies the rater training may be crucial to stabilize the scores across responses.

Educational Implications

Test utility is of great interest in developing performance-based achievement tests. Because performance-based achievement tests can reveal multiple aspects of learning levels, implementation of such tests can support dynamic instructional programs in schools. The National Council of Teachers of Mathematics (NCTM) has stressed fostering problem solving, reasoning, and communication in mathematics education. Assessment should seek the evidence of reasoning processes in solving problems. Communication is the vehicle by which students can appreciate mathematics as the processes of problem solving and reasoning (NCTM, 1991, p.96). Format of tasks in assessment is also an important factor affecting students' performances. Although open-ended questions are more language-dependent than multiple-choice questions, open-ended questions can offer more insight into students' thought than multiple-choice tests (NCTM, 1995).

Fostering reasoning and communication skills are not easy processes for both students and teachers. Many teachers often demonstrate how to communicate mathematically in lessons. However, if tests are not consistent with instructional goals, it is natural that students will not understand what they are expected in mathematics learning. Most students tend to believe that finding a correct answer is the goal of solving math problems.

Examples of students' responses can be used as effective teaching materials for lessons, because they are good examples of how to communicate mathematically or how to improve communications in written formats. Also, multidimensional measurements in scoring rubrics, such as analytical scoring methods, can be indicators to assess cognitive growth for individual student. Therefore, large-scale, performance-based achievement testing can be a good device to measure each student's learning. Although there exist many technical concerns about performance-based achievement tests, they can measure different aspects of learning which have not been measured in multiple-choice type tests. Ranking school is not of among interest for this type of achievement test, but fostering students' cognitive growth is of great interest for developing these types of tests.

Conclusions

Generalizability analyses demonstrated that a minimum number of items on a test can be five items if well chosen. When nine items are contained in a test, the dependability coefficients can exceed more than .9 level. An indicator of "well chosen" items is point-biserial coefficients, which indicate item-test consistency. The magnitude of item-test consistency may depend more on students' performances than task structure. Because statistics such as point-biserial coefficients, the generalizability and dependability coefficients are affected by students' performances, low values of statistical evidence can be caused by low students' performance levels. That is, the higher students' performances become, the higher the statistical evidence may attain. The criteria of "well chosen" items should be examined not only statistically but also qualitatively. Statistical analyses of the tests demonstrates characteristics of students learning at each school, which are useful indicators for examining local educational practices Performance-based tasks can be developed from multiple-choice items.

References

Ackerman, T. A. (1994). LIKERT [Computer software]. Champaign, IL: University of Illinois.

Brennan, R., L. (1983). Elements of generalizability theory. Iowa City, IA: The American College Testing Program

Crick. J. E., & Brennan, R. L. (1983) A general purpose analysis of variance system (GENOVA) (Version 2.2) [Computer software]. Iowa City, IA: American College Testing Program.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability of scores and profiles. New York: John Wiley.

Dunbar, S. B., Koretz, D. M., & Hoover H. D. (1991). Quality control in the development and use of performance assessments. Applied Measurement in Education, 4(4), 289 - 304.

Harnisch, D. L. (1994). Performance assessment in review: New directions for assessing student understanding. International Journal of Educational Research, 21(3), 341-350.

Lane, S. (1993). The conceptual framework for the development of a mathematics performance assessment, Educational Measurement: Issues and Practice, 12(2), 16-23.

Lane, S., Liu, M., Stone, C. A., & Ankenmann, R. D. (1993, April). Validity evidence for QUASAR's mathematics performance assessment. Paper presented at the annual meeting of the American Educational Research Association; Atlanta, GA.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1994). Reliability and validity of a mathematics performance assessment. International Journal of Educational Research, 21(3), 247 - 266.

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

Miller, M. D., & Crocker, L. (1990). Validation methods for direct writing assessment. Applied Measurement in Education, 3(3), 285 - 296.

National Council of Teachers of Mathematics. (1991). Professional standards for teaching mathematics. Reston, VA: Author.

National Council of Teachers of Mathematics. (1995). Assessment standards for school mathematics. Reston, VA: Author.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher, 21(4), 22 - 27.

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage.

Silver, E. A. (1991). Quantitative understanding: Amplifying student achievement and reasoning. Pittsburgh, PA: Learning Research and Development Center.

Suzuki, K., & Harnisch, D. L. (1996, April). Measuring levels of mathematical thinking in algebra: Do test formats affect students' thinking? Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Tatsuoka, K. K. (1992). [A list of attributes for SAT mathematics]. Unpublished working manuscript.

Table 1

Item Composition by Form

	ITEM							
FORM	1	2	3	4	5	6	7	8
A	○	○	–	○	–	○	–	○
B	○	–	○	○	–	–	○	○
C	○	○	○	–	○	○	–	–
D	–	○	○	-	○	-	○	○

Note. ○ indicates presence of the item in the form.
– indicates absence of the item in the form.

Table 2

Ability Measured by Item

ITEM	ABILITY MEASURED
1	• considering two variables having two conditions simultaneously • transforming two equations in two variables into a quadratic equation in one variable
2	• working through two variables by finding conditions from the verbal expressions • solving pair of linear equations
3	• finding two consecutive integers having a condition through word problem. • solving a quadratic equation in one variable
4	• understanding the meaning and relation of a square root and a power of a number
5	• handling two variables having two conditions: an inequality and a quadratic relation. • solving two linear equations in three variables
6	• handling three variables with two conditions given in a verbal form and in a mathematical formula
7	• solving a pair of linear equations in two variables
8	• justifying procedures of solving equations with fractional coefficients

Table 3

Attributes' Chart by Item

	TATSUOKA'S ATTRIBUTES												
ITEM	1	2	3	4	5	6	7	8	9	10	11	12	13
1	○	○	○	–	–	–	–	○	○	–	–	–	–
2	○	○	–	–	○	–	–	–	–	○	–	–	–
3	○	○	○	–	○	–	–	○	○	–	–	–	○
4	–	–	○	–	–	–	–	○	○	–	–	–	○
5	–	–	○	–	○	–	○	–	–	–	–	–	○
6	○	○	–	–	○	–	○	○	–	–	–	–	–
7	○	○	–	–	○	–	○	○	–	–	–	–	–
8	–	○	–	–	–	–	○	○	–	–	○	–	–

Note. ○ indicates the presence of the attribute in the item.
– indicates the absence of the attribute in the item.

Tatsuoka's Attributes:

1: Arithmetic
2: Elementary Algebra
3: Advanced Algebra
4: Geometry & Analytic Geometry
5: Word Problems
6: Comparison Format
7: Recall & Understand Simple Computation
8: Application of Rules & Algorithms
9: Selection & Application of Rules & Theorems
10: Reasoning & Logical Thinking
11: Analytical Thinking & Cognitive Restructuring
12: Reading Comprehension
13: Practical, Spontaneous Wisdom
14: Degree of Complexity

Table 4

Attributes Measured by Form

	TATSUOKA'S ATTRIBUTES												
FORM (ITEM)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
A (12468)	3	4	2	0	2	0	2	4	2	1	1	0	1
B (13478)	3	4	3	0	2	0	1	5	3	0	1	0	2
C (12356)	4	4	3	0	4	0	2	3	2	1	0	0	2
D (23578)	3	4	2	0	4	0	3	3	1	1	1	0	2

Note. Scores in each cell represent the total sum of the presence of an attribute for items by form.
"○ " = 1 point, "–" = 0 point.

Table 5

Participants Distribution by Form by School

		FORM				
COUNTRY	**SCHOOL**	**A**	**B**	**C**	**D**	**TOTAL**
U.S.	**1**	12	12	12	11	47
U.S.	**2**	26	21	26	22	95
JAPAN	**3**	36	37	36	39	148
	TOTAL	**74**	**70**	**74**	**72**	**290**

Table 6

Mean Total Score by School

SCHOOL	n	Mean	Post-hoc
1	47	11.85	↕ * ↕ * ↕ *
2	95	13.42	
3	148	16.19	

Note. * stands for significant difference between two groups in Scheffe's post-hoc test.

Table 7

Mean Total Score by Form by School

	FORM			
SCHOOL	A	B	C	D
1	11.42	10.67	14.92	10.27
2	14.42	12.57	13.58	12.86
3	16.03	17.43	15.31	15.97

Table 8

Significant Different Items on Mean Scores at School Levels

	ITEM							
ITEM	1	2	3	4	5	6	7	8
1	—		S2		S2, S3		S1, S3	S1
2		—	S2, S3	S2	S2, S3	S2	S2, S3	S2
3			—	S1	S1, S3	S3	S1	S1
4				—	S2, S3	S3	S2	
5					—	S3	S3	S2, S3
6						—	S3	
7							—	
8								—

Note. Each entry in each cell indicates the significant difference in mean scores between the two items.
S1 stands for School 1, S2 stands for School 2, and S3 stands for School 3.

Table 9

Significant Correlations Among Items at School Levels

	ITEM							
ITEM	1	2	3	4	5	6	7	8
1	—	S2*	S1**, S2**, S3*				S2*	
2		—	S1*, S2**, S3**		S3**	S1*, S2*	S1**, S3**	S2*, S3*
3			—		S2*, S3**	S3*	S3**	S3*
4				—				
5					—	S2*	S3*	S3**
6						—		
7							—	S3**
8								—

Note. Each entry in each cell indicates the significant difference in mean scores between the two items.
S1 stands for School 1, S2 stands for School 2, and S3 stands for School 3.
* p < .05, ** p < .01

Table 10

Coefficient Alphas by Form by School

FORM	SCHOOL	MEAN *	SD	ALPHA
A	1	11.42	3.15	.00
	2	14.42	2.90	.49
	3	16.03	2.34	.30
B	1	10.67	3.25	.31
	2	12.57	3.33	.66
	3	17.43	2.61	.55
C	1	14.92	2.84	.53
	2	13.58	3.08	.65
	3	15.31	2.90	.45
D	1	10.27	2.93	.22
	2	12.86	2.05	.33
	3	15.97	3.24	.78

Note. * Each form has 20 points worth in total scores.

Table 11

Variance Estimates for the Person x Item Generalizability Studies Given Random Effects Models.

FORM A

	SCHOOL 1		SCHOOL 2		SCHOOL 3	
	variance	%	variance	%	variance	%
Person	—	—	.17	14	.07	7
Item	—	—	.12	10	.08	9
P x I	2.39	—	.89	76	.79	84

FORM B

	SCHOOL 1		SCHOOL 2		SCHOOL 3	
	variance	%	variance	%	variance	%
Person	.14	7	.31	28	.15	19
Item	.28	14	.01	0	.02	2
P x I	1.59	79	.80	72	.63	79

FORM C

	SCHOOL 1		SCHOOL 2		SCHOOL 3	
	variance	%	variance	%	variance	%
Person	.19	13	.26	21	.16	12
Item	.37	26	.29	23	.18	14
P x I	.83	61	.69	56	.95	74

FORM D

	SCHOOL 1		SCHOOL 2		SCHOOL 3	
	variance	%	variance	%	variance	%
Person	.08	2	.06	6	.33	30
Item	.26	14	.37	36	.31	28
P x I	1.47	84	.59	58	.48	42

Table 12

Generalizability and Dependability Coefficients for Person x Item (P x I) Design and Item nested in Person (I : P) Design Decision Studies

FORM A

		(P x I) design			(I : P) design		
	n'_i	SCHOOL 1	SCHOOL 2	SCHOOL 3	SCHOOL 1	SCHOOL 2	SCHOOL 3
ρ^2	5	–	.49	.30	–	.46	.28
	7	–	.57	.37	–	.54	.35
	9	–	.63	.43	–	.60	.41
ϕ	5	–	.46	.28	–	.46	.28
	7	–	.54	.35	–	.54	.35
	9	–	.60	.41	–	.60	.41

Note. ρ^2 stands for a generalizability coefficient. ϕ stands for a dependability coefficient.

FORM B

		(P x I) design			(I : P) design		
	n'_i	SCHOOL 1	SCHOOL 2	SCHOOL 3	SCHOOL 1	SCHOOL 2	SCHOOL 3
ρ^2	5	.31	.66	.55	.28	.66	.54
	7	.39	.73	.63	.35	.73	.62
	9	.45	.78	.69	.41	.77	.68
ϕ	5	.28	.66	.54	.28	.66	.54
	7	.35	.73	.62	.35	.73	.62
	9	.41	.77	.68	.41	.77	.68

FORM C

		(P x I) design			(I : P) design		
	n'_i	SCHOOL 1	SCHOOL 2	SCHOOL 3	SCHOOL 1	SCHOOL 2	SCHOOL 3
ρ^2	5	.53	.65	.45	.44	.57	.41
	7	.61	.72	.53	.52	.65	.49
	9	.67	.77	.60	.58	.70	.55
ϕ	5	.44	.57	.41	.44	.57	.41
	7	.52	.65	.49	.52	.65	.49
	9	.58	.70	.55	.58	.70	.55

FORM D

		(P x I) design			(I : P) design		
	n'_i	SCHOOL 1	SCHOOL 2	SCHOOL 3	SCHOOL 1	SCHOOL 2	SCHOOL 3
ρ^2	5	.22	.33	.78	.19	.23	.68
	7	.29	.41	.83	.25	.30	.75
	9	.34	.47	.86	.30	.35	.79
ϕ	5	.19	.23	.68	.19	.23	.68
	7	.25	.30	.75	.25	.30	.75
	9	.30	.35	.79	.30	.35	.79

Table 13

Descriptive Statistics for Selected Items' Form

FORM	SCHOOL	ITEMS	MEAN*	SD	ALPHA
A	1	4, 6, 8	6.00	2.92	.41
	2	1, 2, 6	8.50	2.36	.55
	3	1, 6, 8	9.31	2.07	.31
B	1	1, 3, 4	7.58	3.30	.79
	2	1, 3, 7	7.48	2.65	.78
	3	1, 3, 8	10.32	2.01	.69
C	1	1, 5, 6	8.33	2.49	.65
	2	3, 5, 6	7.23	2.56	.67
	3	3, 5, 6	8.89	2.54	.60
D	1	2, 3, 7	6.82	3.13	.76
	2	2, 3, 8	8.82	1.80	.67
	3	2, 3, 8	9.92	2.24	.72**

Note. * Each selected items' form has 12 points worth in total scores.
** This alpha value was less than the original 5 item form.

Table 14

Variance Estimates for the Person x Item Generalizability Studies Given Random Effects Models for Selected Items' Form.

SELECTED ITEMS' FORM A

	A468*		A126		A168	
	SCHOOL 1		SCHOOL 2		SCHOOL 3	
	variance	%	variance	%	variance	%
Person	.42	19	.35	24	.15	12
Item	(0)	0	.26	17	.05	4
P x I	1.82	81	.87	59	1.01	84

Note. * The letter A stands for a form type followed by the numbers indicating selected items in the form. e.g. A468 stands for item 4, 6, 8, selected in Form A.

SELECTED ITEMS' FORM B

	B134		B137		B138	
	SCHOOL 1		SCHOOL 2		SCHOOL 3	
	variance	%	variance	%	variance	%
Person	1.05	52	.64	54	.32	41
Item	.16	8	(0)	0	.01	2
P x I	.82	40	.54	46	.44	57

SELECTED ITEMS' FORM C

	C156		C356		C356	
	SCHOOL 1		SCHOOL 2		SCHOOL 3	
	variance	%	variance	%	variance	%
Person	.49	28	.51	39	.44	26
Item	.44	26	.05	4	.35	21
P x I	.79	56	.75	57	.88	53

SELECTED ITEMS' FORM D

	D124		D238		D238	
	SCHOOL 1		SCHOOL 2		SCHOOL 3	
	variance	%	variance	%	variance	%
Person	.91	40	.25	28	.41	43
Item	.52	23	.26	30	.07	7
P x I	.87	37	.37	42	.47	50

Table 15

Generalizability and Dependability Coefficients for Person x Item (P x I) Design and Item nested in Person (I : P) Design Decision Studies for Selected Items' Form

SELECTED ITEMS' FORM A

		(P x I) design			(I : P) design		
		A468*	A126	A168	A468	A126	A168
	n_i'	SCHOOL 1	SCHOOL 2	SCHOOL 3	SCHOOL 1	SCHOOL 2	SCHOOL 3
ρ^2	5	.54	.67	.42	.54	.61	.41
	7	.62	.74	.51	.62	.69	.50
	9	.68	.78	.57	.68	.74	.56
ϕ	5	.54	.61	.41	.54	.61	.41
	7	.62	.69	.50	.62	.69	.50
	9	.68	.74	.56	.68	.74	.56

Note. ρ^2 stands for a generalizability coefficient. ϕ stands for a dependability coefficient.
* The letter A stands for a form type followed by the numbers indicating selected items in the form. e.g. A468 stands for item 4, 6, 8, selected in Form A.

SELECTED ITEMS' FORM B

		(P x I) design			(I : P) design		
		B134	B137	B138	B134	B137	B138
	n_i'	SCHOOL 1	SCHOOL 2	SCHOOL 3	SCHOOL 1	SCHOOL 2	SCHOOL 3
ρ^2	5	.87	.85	.78	.84	.85	.78
	7	.90	.89	.84	.88	.89	.83
	9	.92	.91	.87	.91	.91	.86
ϕ	5	.84	.85	.78	.84	.85	.78
	7	.88	.89	.83	.88	.89	.83
	9	.91	.91	.86	.91	.91	.86

SELECTED ITEMS' FORM C

		(P x I) design			(I : P) design		
		C156	C356	C356	C156	C356	C356
	n_i'	SCHOOL 1	SCHOOL 2	SCHOOL 3	SCHOOL 1	SCHOOL 2	SCHOOL 3
ρ^2	5	.76	.77	.71	.67	.76	.64
	7	.81	.82	.78	.74	.81	.71
	9	.85	.86	.82	.78	.85	.76
ϕ	5	.67	.76	.64	.67	.76	.64
	7	.74	.81	.71	.74	.81	.71
	9	.78	.85	.76	.78	.85	.76

SELECTED ITEMS' FORM D

		(P x I) design			(I : P) design		
		D124	D238	D238	D124	D238	D238
	n_i'	SCHOOL 1	SCHOOL 2	SCHOOL 3	SCHOOL 1	SCHOOL 2	SCHOOL 3
ρ^2	5	.84	.77	.81	.76	.67	.79
	7	.88	.83	.86	.82	.74	.84
	9	.90	.86	.89	.85	.78	.87
ϕ	5	.76	.67	.79	.76	.67	.79
	7	.82	.74	.84	.82	.74	.84
	9	.85	.78	.87	.85	.78	.87

Table 16

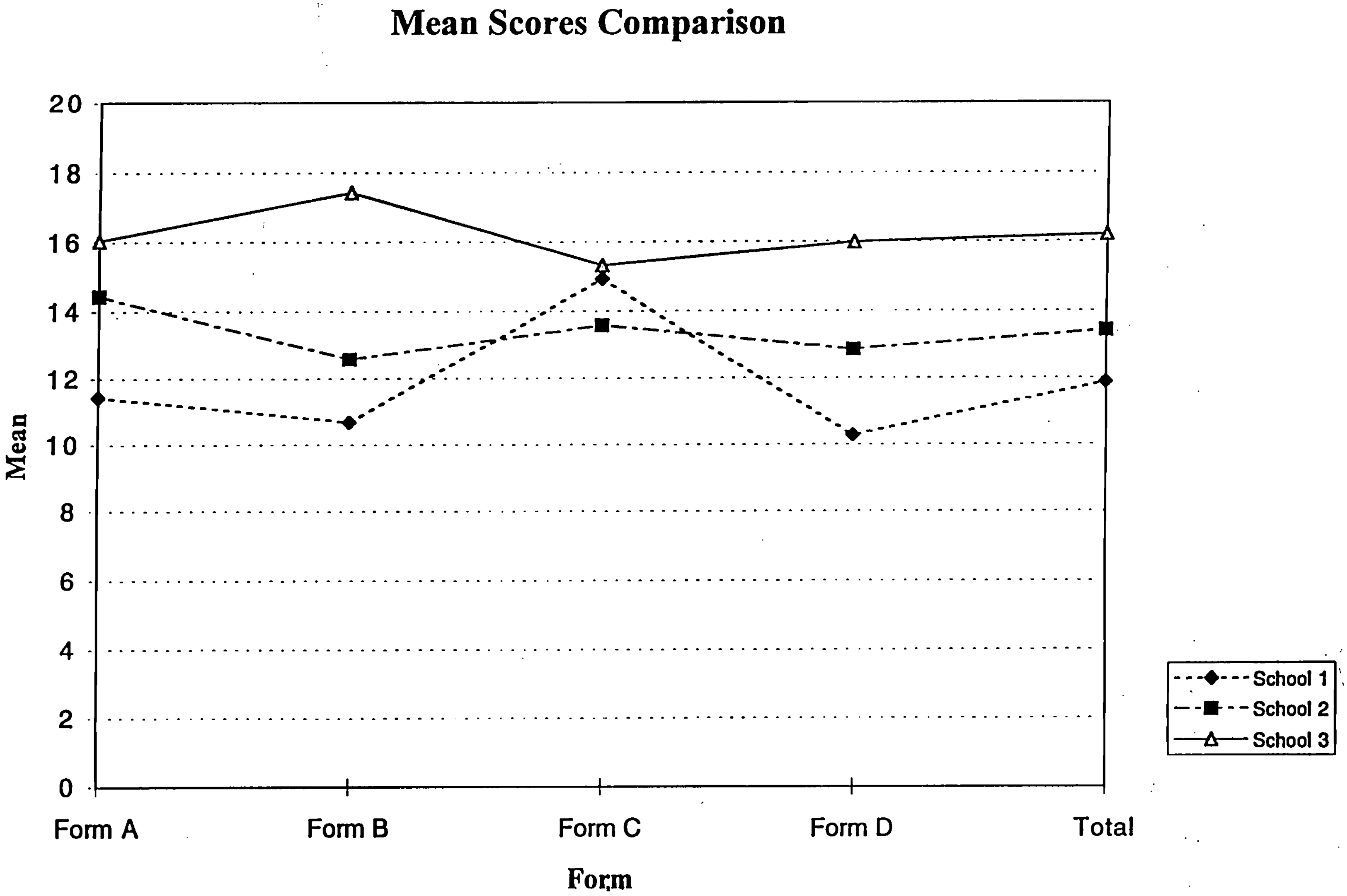
Item Selection by Form by School

		ITEM							
FORM	SCHOOL	1	2	3	4	5	6	7	8
A	1			▨	○	▨	○	▨	○
	2	○	○	▨		▨	○	▨	
	3	○		▨		▨	○	▨	○
B	1	○	▨	○	○	▨	▨		
	2	○	▨	○		▨	▨	○	
	3	○	▨	○		▨	▨		○
C	1	○			▨	○	○	▨	▨
	2			○	▨	○	○	▨	▨
	3			○	▨	○	○	▨	▨
D	1	▨	○	○	▨		▨	○	
	2	▨	○	○	▨		▨		○
	3	▨	○	○	▨		▨		○

Note. Dark shaded area shows no inclusion of the items in the form.

○ indicates the item was selected for selected items' form because of its higher point-biserial coefficient.

Figure 1


Mean Scores Comparison
20
18
16
14
12
10
8
6
4
2
0
Mean
Form A
Form B
Form C
Form D
Total
Form
School 1
School 2
School 3

A List of Attributes for SAT Mathematics

Attribute	Description
1. Arithmetic	Can recall and apply knowledge of basic properties and operations in arithmetic. Students also understand the meaning of basic concepts and properties in arithmetic. These are, for example, knowing numbers which include: whole numbers, fractions, decimals and sign-numbers, and their differences and simple sub-features such as odd and even numbers, prime numbers and factors. The basic operations include addition, subtraction, multiplication, and division. The basic concepts include: percentages, ratios, powers, roots, and logarithms for numbers (not including variables). The deeper understanding of the concepts and their interrelationships are not included. Students are expected to know the simple and basic concept of ordering, and can order numbers and represent the order on the number line.
2. Elementary Algebra	Can recall and apply knowledge of basic properties and basic operations in algebra such as addition and subtraction of numbers and variables. Students should also understand the meaning of basic concepts such as variables, linear equations and simple (linear) algebraic expressions.
3. Advanced Algebra	Can recall knowledge of properties and operations in algebra. The difference between elementary and advanced algebra is that the former deals with the variables with the first degree only, and hence only linear expressions of equations are included. The advanced algebra covers higher degree expressions such as quadratic equations, polynomial expressions and more complicated algebraic expressions. This attribute includes knowledge and understanding of simple concepts of functions, probability, and combinatorics.
4. Geometry & Analytic Geometry	Can recall knowledge of geometric figures and understand perimeters, areas, and volumes for triangles, circles, rectangles, and other geometric objects. Students also understand points, lines, planes and their expressions in terms of coordinates.
5. Word Problems	Can transform a verbal representation into an algebraic expression or equation with one or two unknown variables in which the transformation is straightforward and does not require reasoning or logical thinking.
6. Comparison Format	Item format is a comparison type.
7. Recall & Understand Simple Computations	Recall knowledge and understand the meaning of definitions, concepts, properties, theorems and operational rules and algorithms that are not included in Attributes 1 through 4. This attribute includes understanding of relationships among the topics above. Simple computation using basic operations stated in Attributes 1 through 3 are included.
8. Application of Rules & Algorithms	Can solve equations, derive algebraic expressions, factor algebraic equations and expressions, and substitute successfully the numbers and/or variables in algebraic expressions and equations.
9. Selection & Application of Rules & Theorems	Can select a relevant concept, or a theorem, or rule for solving the problem of interest from a large domain of knowledge space. Then can apply it correctly.
10. Reasoning & Logical Thinking	Can create an equation or simultaneous equations from a word problem that cannot be translated easily into an algebraic expression or equation. Can solve inequalities and can check necessary and sufficient conditions for solving equations or problems. Can solve a problem deductively.
11. Analytical Thinking & Cognitive Restructuring	Can do analysis for separating something (implicit) into component parts or constituent elements, and restructure it cognitively so that a problem becomes solvable. Requires higher mental processes.
12. Reading Comprehension	Can read complex verbal problems and can follow instructions. The instructions include mathematical and geometrical terminology. Can relate the geometric terms with figures, graphs, and charts.
13. Practical, Spontaneous Wisdom	Can solve a problem by going back from the options, or by making one or two examples and infer the correct option. Sometimes, the problem is solved intuitively but reasoning is very unconventional. Substituting numbers is often used for selecting the correct option.
14. Degree of complexity	Can sort goals and subgoals (explicit) embedded in a problem. Can solve several goals step by step.

Appendix B - (1)

Eight Items

1. When $xy = 60$ and $x + y = 17$, find the value of $x - y$. Show all your work and explain in words how you found your answer.

2. In an election, a total of 620 people each voted once for one of the candidates, as shown in the table below. If twice as many people voted for Clinton as for Smith, how many voted for Clinton? Show all your work and explain in words how you found your answer.

CANDIDATE	NUMBER OF VOTES
Kennedy	280
Clinton	x
Smith	y
Bush	70

3. Sara opened her book and realized that when she multiplied the page numbers she saw, the product equaled 272. To which pages did she open? Write an equation to represent the given relationship among the integers, then solve it. Show all your work and explain in words how you found your answer.

4. If x can be either 2 or 8, and if n can be any integer from 1 through 10, inclusive, for how many different combinations of x and n will $(\sqrt{x})^n$ be an integer? Show all your work and explain in words how you found your answer.

5. If x and y are integers and $y < 20$, for exactly how many different ordered pair (x, y) will $x^2 = y$? List all pairs you found. Show all your work and explain in words how you found your answer.

6. If X, Y, and Z are different odd digits in the correctly worked sum of three two-digit numbers shown below, find the value of Y.

$$XY + XY + XY = YZ$$

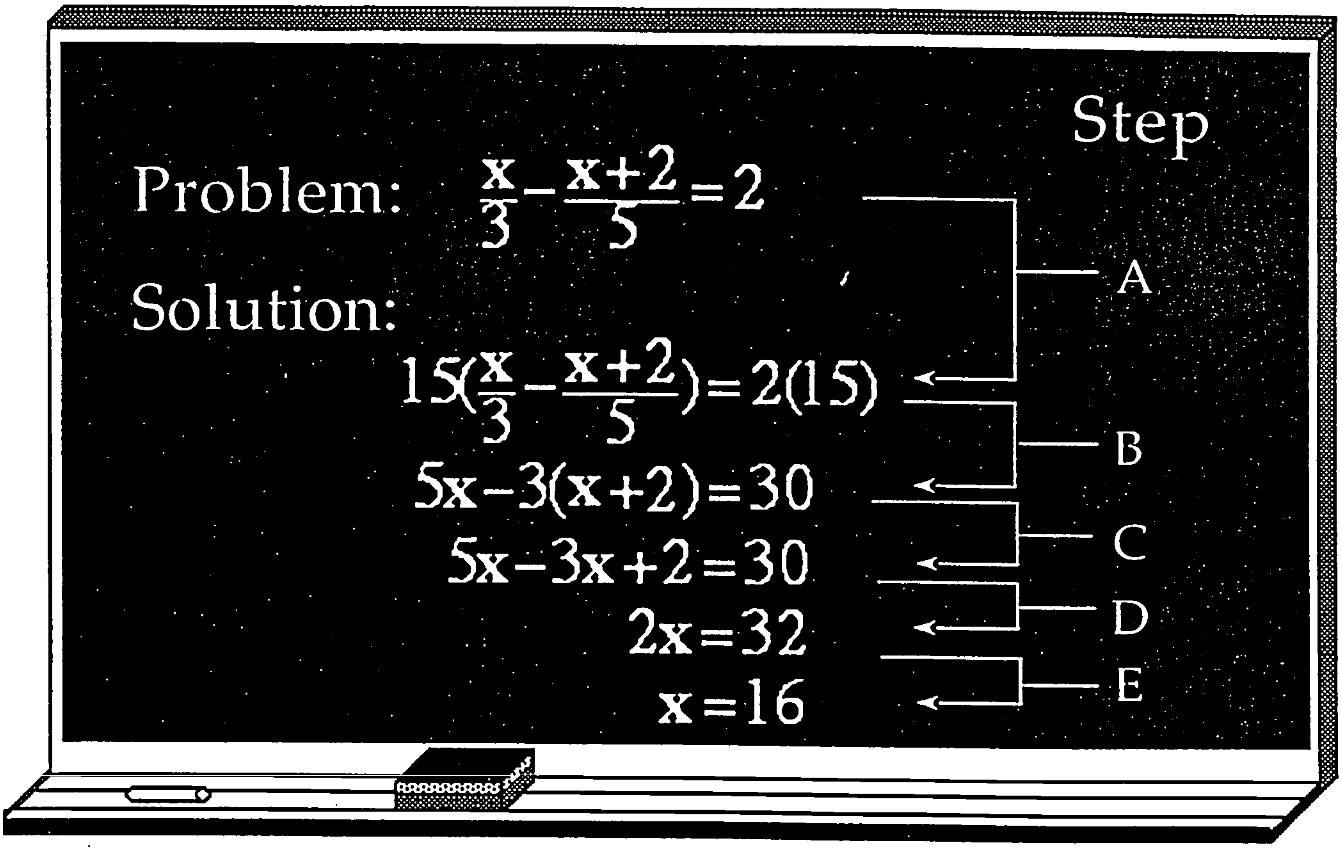
Show all your work and explain in words how you found your answer.

7. John worked at a fast-food restaurant. He was paid \$5 per hour Monday through Saturday. On Sunday he was paid time and a half (one and one half times his hourly pay). He worked 45 hours in a week and had a total weekly pay of \$250. How many hours did he work on Sunday? Show all your work and explain in words how you found your answer.

Appendix B - (2)

8. One day in algebra class, Susie raised her hand and said,
"I solved my homework yesterday, but I think something is wrong in my solution. When I substitute my answer to the left side of the original equation, it doesn't equal the value on the right side. However, I can't figure out what's wrong with my solution. Will you tell me why my solution is not right?"

Her teacher asked her to write her answer on the blackboard.


Step
Problem: $\frac{x}{3}-\frac{x+2}{5}=2$
A
Solution:
$15(\frac{x}{3}-\frac{x+2}{5})=2(15)$
B
$5x-3(x+2)=30$
C
$5x-3x+2=30$
D
$2x=32$
E
$x=16$

Her teacher wrote down A, B, C, D, and E for each step.
Now you are asked to evaluate each of the steps.
Explain why the step is correct or incorrect.

< Continued on following page>

Appendix B - (3)

For each of the steps: if correct, justify why it's correct, if incorrect, explain why.

STEP	CORRECT/INCORRECT	JUSTIFICATION
A		
B		
C		
D		
E		

MATHEMATICS SCORING RUBRIC: A GUIDE TO SCORING OPEN-ENDED ITEMS

SCORE LEVEL	MATHEMATICAL KNOWLEDGE Knowledge of mathematical principles and concepts which result in a correct solution to a problem	STRATEGIC KNOWLEDGE Identification of important elements of the problem and the use of models, diagrams and symbols to systematically represent and integrate concepts	COMMUNICATION Written explanation and rationale for the solution process
4	• shows complete understanding of the problem's mathematical concepts & principles • uses appropriate mathematical terminology & notation (e.g. labels answer as appropriate)[1] • executes algorithms completely and correctly	• identifies all the important elements of the problem and shows complete understanding of the relationships among elements • reflects an appropriate and systematic strategy for solving the problem • gives clear evidence of a complete and systematic solution process	• gives a complete written explanation of the solution process employed; explanation addresses what was done, and <u>why</u> it was done • if a diagram is appropriate, there is a complete explanation of all the elements in the diagram
3	• shows nearly complete understanding of the problem's mathematical concepts and principles • uses nearly correct mathematical terminology and notations • executes algorithms completely; computations are generally correct but may contain minor errors	• identifies most of the important elements of the problem and shows general understanding of the relationships among them • reflects an appropriate strategy for solving the problem • solution process is nearly complete	• gives a nearly complete written explanation of the solution process employed; may contain some minor gaps • may include a diagram with most of the elements explained
2	• shows some understanding of the problem's mathematical concepts and principles • may contain major computational errors	• identifies some important elements of the problem but shows only limited understanding of the relationships among them • appears to reflect an appropriate strategy but application of strategy is unclear • gives some evidence of a solution process	• gives some explanation of the solution process employed, but communication is vague or difficult to interpret • may include a diagram with some of the elements explained
1	• shows limited to no understanding of the problem's mathematical concepts and principles • may misuse or fail to use mathematical terms • may contain major computational errors	• fails to identify important elements or places too much emphasis on unimportant elements • may reflect an inappropriate strategy for solving the problem • gives minimal evidence of a solution process; process may be difficult to identify • may attempt to use irrelevant outside information	• provides minimal explanation of solution process; may fail to explain or may omit significant parts of the problem • explanation does not match presented solution process • may include minimal discussion of elements in diagram; explanation of significant elements is unclear
0	• no answer attempted	• no apparent strategy	• no written explanation of the solution process is provided

Appendix D

Scoring Example

Example 1: This student was scored a "4", although the final answer was incorrect. The solution strategy and process were explained well, and the response exhibited complete understanding of the problem. We can infer that this student answered incorrectly because the final answer appeared in mind during processing the response.

Item 1. When xy = 60 and X + Y= 17, find the value of x - y. Show all your work and explain in words <u>how</u> you found your answer.

x	y
1	16
2	15
3	14
4	13
5	12
6	11
7	10
8	9

$$\begin{array}{r} 12 \\ -7 \\ \hline \end{array} \qquad \begin{array}{r} 7 \\ -12 \\ \hline \end{array}$$

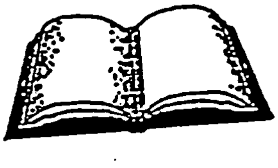
$\boxed{5 \text{ or } -5}$

made a graph showing the values of the sum of the #s. leading to 17. Then I took those sums and multiplied the products till I found a sum that equaled 60. Then I had 12 and 7. Then I took 12-7 and got 5, then 5-12 and got -5. Those are the 2 answers I got.

Example 2: This student was scored a "2", although the correct answer was found. The solution process was not explained in words at all, and the final answer was not determined.

Item 3. Sara opened her book and realized that when she multiplied the page numbers she saw, the product equaled 272. To which pages did she open? Write an equation to represent the given relationship among the integers, then solve it. Show all your work and explain in words <u>how</u> you found your answer.

12×12 = 144
13×14 = 182
14×15 = 210
15×16 = 240
16×17 = 272.



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM030224

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: An Investigation on the Generalizability of Performance-based Assessment in Mathematics	
Author(s): Kyoko Suzuki, Delwyn L. Harnisch	
Corporate Source: University of Illinois at Urbana-Champaign	Publication Date: 4/8/96

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	Level 2A	Level 2B
☑	□	□
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproductio n by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature:	Printed Name/Position/Title: Kyoko Suzuki / Postdoctoral fellow	
Organization/Address: ETS / MS 02-T Educational Testing Service Princeton, NJ 08541	Telephone: 609-734-5079	FAX: 609-734-5420
	E-Mail Address: ksuzuki@ets.org	Date: 9/20/99

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

The Catholic University of America
ERIC Clearinghouse on Assessment and Evaluation
210 O'Boyle Hall
Washington, DC 20064
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com