Effects of Response Format on Diagnostic Assessment of Scholastic Achievement

Menucha Birenbaum, Tel Aviv University

Kikumi K. Tatsuoka, Educational Testing Service

Yaffa Gutvirtz, High School Ironi Daled, Tel Aviv, Israel

The effect of response format on diagnostic assessment of students' performance on an algebra test was investigated. Two sets of parallel, open-ended (OE) items and a set of multiple-choice (MC) items—which were stem-equivalent to one of the OE item sets—were compared using two diagnostic approaches: a "bug" analysis and a rule-space analysis. Items with identical format (parallel OE items) were more similar than items with different formats (OE vs. MC). *Index terms: bug analysis, diagnostic assessment, free-response, item format, multiple-choice, rule space.*

Response formats of assessment measures vary. There are two broad categories of response—constructed response and choice response—with various types of formats subsumed under each. In a constructed-response (also known as free-response) format, an examinee is required to generate an answer to an open-ended (OE) item; in a choice-response format, an examinee is required to select one or more answers from a short list of options. The most common item type in this category is the multiple-choice (MC) item.

Numerous studies have compared the two formats with respect to different domains and from different perspectives. (For recent reviews of the literature see Bennett, 1991; Traub & MacRury, 1990.) Response format comparisons have included theoretical considerations of the cognitive processing requirements of the two formats; empirical investigations concerning the psychometric properties of the two formats; examination of interaction effects of factors such as gender, race/ethnicity, test anxiety, test wiseness, and examinees' attitudes toward the formats; and examination of the effects of format expectancy on test preparation and test performance.

Some of the psychometric properties on which the two response formats have been extensively contrasted include item difficulty and test reliability (e.g., Martinez, 1991; Oosterhof & Coats, 1984; Traub & MacRury, 1990); construct validity (e.g., Bennett, Rock, Braun, Frye, Spohrer, & Soloway, 1990; Van den Bergh, 1990; Ward, 1982; Ward, Frederiksen, & Carlson, 1980); and predictive validity (e.g., Bridgeman, 1991; Bridgeman & Lewis, 1991). Despite strong assertions by cognitive theorists regarding the differences between the cognitive demands of the two formats, the empirical studies have yielded only equivocal evidence for format effects (Traub & MacRury, 1990).

There is little research on the effect of response formats on the diagnostic assessment of scholastic achievement. However, Birenbaum & Tatsuoka (1987) compared OE and MC items in an arithmetic procedural task with respect to various criteria, including the average number of different error types per examinee and the diagnosed sources of misconception. The results indicated considerable differences between the two response formats, favoring the OE format for diagnostic assessment.

The present study further examined the effect

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 4, December 1992, pp. 353-363

0146-6216/92/040353-11$1.80

of response format using additional diagnostic assessment criteria. In this study, comparisons were made between parallel-stem items with identical response format (OE) and stem-equivalent items with different response formats (OE vs. MC). In addition to comparing the diagnostic results of different formats, parallel OE items were contrasted to address the issue of "bug" instability (see, e.g., Payne & Squibb, 1990; Sleeman, Kelly, Martinak, Ward, & Moore, 1989). A bug (or "mal-rule") is an incorrect rule that an examinee uses to solve a problem. To the extent that bugs are unstable, the diagnostic results of subsets with the same format or different formats may be affected. The criteria for examining the response format effect were: (1) the extent of similarity between the diagnostic results from each of the three subsets, and (2) the percent of matched bugs and matched sources of errors (task attributes) in the three subsets. The procedural task used was the solution of algebraic linear equations with one unknown. Two diagnostic approaches were employed: A bug analysis (a deterministic approach for identifying the mal-rules underlying the examinees' response patterns; e.g., Brown & Burton, 1978), and a probabilistic approach—rule space analysis (Tatsuoka, 1983, 1985, 1990, 1991; Tatsuoka & Tatsuoka, 1987).

Method

Examinees

The sample consisted of 231 8th and 9th graders (ages 14–15) from a high school in Tel Aviv that was heterogeneous with respect to math ability and achievement. 57% of the examinees were girls. The 8th and 9th graders had been grouped into high ($N = 106$) and low ($N = 125$) mathematics achievement groups.

Instruments and Procedures

A 48-item diagnostic test consisting of linear algebraic equations with one unknown was developed by Gutvirtz (1989), based on a detailed task analysis that included a procedural network and a mapping sentence (e.g., Birenbaum & Shaw, 1985). The test was developed to identify students' bugs in solving linear algebraic equations. 32 OE items were used, and students were asked to show all work toward the solution. For the 16 MC items used, five or six response options representing frequent errors were provided. The MC items followed the OE items. The instructions for the MC items did not mention guessing. To prevent cheating, two forms of the algebra test were constructed—Form I and Form II. Each form contained Subsets 1 and 2 (OE items) that were parallel in their attributes and Subset 3 (MC items) that had stems identical to the stems of the items in Subset 1. Table 1 shows the three subsets for Form I (8 items per subset) and Form II (7 items per subset). The item response theory and classical test theory difficulty and discrimination indices and Cronbach's α reliability coefficients of the subsets also are provided in Table 1.

The Bug Analysis

Based on a detailed examination of the procedures used by the students to solve the test items, 35 mal-rules (bugs) were identified. The following are examples of five mal-rules used by students to solve the item $ax = b$:

1. $x = a + b$
2. $x = a - b$
3. $x = b - a$
4. $x = -(a + b)$
5. $x = b$

After the bugs were identified, the OE test items were answered systematically by the test developer according to each of the 35 mal-rules. A bug matrix of order $b \times n$ was constructed, where b is the number of bugs, and n is the number of test items. The entries of this matrix were the responses to the test items produced by the mal-rules. The students' actual responses were matched to the entries in the bug matrix and coded accordingly. Of the actual responses to Forms I and II, 94.7% and 94.6%, respectively, were matched to identified bugs or to the correct rule; the rest were either unidentified bugs, clerical errors, or omissions. 38 different response

Table 1
OE and MC Items of Form I and II, Their IRT Difficulty (b) and Discrimination (a) Parameter Estimates, Proportion Correct, (p), and Item-Total Point-Biserial Correlations r_{pbis}, and Cronbach's Alpha Reliabilities (α) of the Subsets

Item Number	Item Stem	IRT Parameters a	b	p	r_{pbis}
Form I, Subset 1: First OE Subset (α = .80)					
1	$4 + x = 6 + 2 \times 3$	1.10	-.89	.75	.59
2	$28x = 7$	.98	-.14	.55	.50
3	$24 = 6x$	1.96	-1.77	.93	.60
4	$8 + 4x = 26$	1.56	-1.30	.86	.62
5	$6(x + 3) = 12x$	1.13	-1.04	.79	.57
6	$5 + 3x + x = 16$	1.16	-.87	.75	.57
7	$75 = 5 + 5x$	1.65	-1.13	.84	.65
8	$x - 6 = 3 + 5 \times 3$	.87	-.59	.66	.49
Form I, Subset 2: Second OE Subset (α = .79)					
1	$13 + x = 6 + 3 \times 2$	.86	-1.02	.75	.51
2	$16x = 4$	.69	-.49	.62	.41
3	$35 = 7x$	1.33	-2.16	.95	.45
4	$3 + 6x = 18$	.81	-1.24	.79	.45
5	$4(2x + 3) = 10x$	1.53	-1.07	.82	.66
6	$6 + 4x + x = 22$	1.93	-.85	.79	.69
7	$98 = 7 + 7x$	1.72	-1.16	.85	.68
8	$x - 4 = 4 + 2 \times 4$	.75	-1.07	.74	.46
Form I, Subset 3: MC Subset (α = .71)					
1	$4 + x = 6 + 2 \times 3$	.82	-.30	.58	.47
2	$28x = 7$	.75	.21	.45	.41
3	$24 = 6x$	.51	-2.96	.91	.27
4	$8 + 4x = 26$	1.42	-1.34	.86	.63
5	$6(x + 3) = 12x$	.44	1.58	.26	.21
6	$5 + 3x + x = 16$	.87	-1.15	.78	.50
7	$75 = 5 + 5x$	1.16	-1.35	.85	.57
8	$x - 6 = 3 + 5 \times 3$	.69	-.73	.67	.44
Form II, Subset 1: First OE Subset (α = .85)					
1	$13 + x = 6 + 3 \times 2$	1.32	-.70	.73	.64
2	$16x = 4$	1.45	-.40	.65	.62
3	$35 = 7x$	1.51	-1.68	.92	.58
4	$3 + 6x = 18$	3.19	-.61	.75	.72
5	$4(2x + 3) = 10x$	2.16	-1.06	.84	.74
6	$6 + 4x + x = 22$	1.72	-.72	.75	.65
7	$x - 4 = 4 + 2 \times 4$	.86	-.85	.72	.52
Form II, Subset 2: Second OE Subset (α = .80)					
1	$4 + x = 6 + 2 \times 3$	1.10	-.70	.71	.60
2	$28x = 7$	.93	-.06	.53	.49
3	$24 = 6x$	1.20	-1.73	.90	.51
4	$8 + 4x = 26$	1.61	-1.10	.83	.65
5	$6(x + 3) = 12x$	1.68	-1.08	.83	.68
6	$5 + 3x + x = 16$	2.28	-.73	.77	.71
7	$x-6 = 3 + 5 \times 3$	.85	-.70	.68	.53

continued on the next page

Table 1, continued
OE and MC Items of Form I and II, Their IRT Difficulty (b) and Discrimination (a) Parameter Estimates, Proportion Correct, (p), and Item-Total Point-Biserial Correlations r_{pbis}, and Cronbach's Alpha Reliabilities (α) of the Subsets

Item Number	Item Stem	IRT Parameters a	b	p	r_{pbis}
Form II, Subset 3: MC Subset (α = .74)					
1	$13 + x = 6 + 3 \times 2$	.83	-.91	.73	.49
2	$16x = 4$	.73	-.61	.65	.43
3	$35 = 7x$	.47	-2.44	.85	.31
4	$3 + 6x = 18$	.93	-1.20	.80	.54
5	$4(2x + 3) = 10x$	.87	-.65	.68	.48
6	$6 + 4x + x = 22$	1.36	-.91	.78	.64
7	$x - 4 = 4 + 2 \times 4$	.87	-1.30	.81	.54

codes were used—one code indicated correct responses, one indicated unidentified errors, one indicated clerical errors, and the rest indicated the 35 identified bugs. The codes for parallel items then were compared. Matches and mismatches were counted across the pairs of parallel items for each student, and classified according to the following primary categories:

A = matched correct (1,1);
B = one correct and one error (1,0; 0,1);
C = matched bug; and
D = unmatched errors (unequal bugs, unidentified bugs, or clerical errors).

The responses to the MC items were coded according to a prespecified key in which each distractor corresponded to a bug. The coded responses were matched to those of the stem-equivalent OE items using the above-mentioned categories.

The Rule-Space Analysis

To conduct a rule-space analysis, the cognitive requirements of the task (also called attributes) are determined. Then an incidence matrix **Q** is constructed; this matrix indicates which attributes are involved in solving each item. **Q** is binary and of order $K \times m$ (the number of attributes $\times$ the number of items). If q_{kj} is the (k, j) element of this matrix (where k indicates an attribute and j indicates an item), then $q_{kj} = 1$ if item j involves attribute k, and $q_{kj} = 0$, otherwise. Cognitive patterns represented by unobservable variables that can be derived from the incidence matrix **Q** are called knowledge states (or cognitive states or attribute patterns). Boolean description functions are used to systematically determine those knowledge states and map them into observable item-score patterns (called ideal item-score patterns; Tatsuoka, 1991; Varadi & Tatsuoka, 1989). It is assumed that an item can be answered correctly if and only if all the attributes associated with the item have been mastered. The knowledge states are represented by a list of mastered/not mastered (or "can/cannot") attributes. The increase of the number of states is combinatorial, but Boolean algebra provides the mathematical tools to overcome the problem of combinatorial explosion.

Once the knowledge states (ideal item-score patterns) are obtained, the actual data are considered. The actual item-score patterns of the students are mapped onto the knowledge states to determine the ideal item-score pattern closest to each student's actual response pattern. This pattern classification problem is handled by the rule-space model, which formulates the classification space and procedures. Item response theory (IRT) is used to formulate the classification space, which is a Cartesian product space of IRT ability/proficiency (θ) and variable(s) ζ

that measure the unusualness (appropriateness/person fit) of item-score patterns (Tatsuoka, 1984; Tatsuoka & Linn, 1983). Bayes' decision rules are used to classify students into knowledge states. Once classified, the attributes a given student is likely to have mastered or failed to master can be indicated.

Application of Rule-Space Analysis

Determining the attributes. A set of nine attributes was specified, which indicated a possible solution strategy for solving the test items (see Table 2). These data were used to produce the incidence matrices.

Testing the adequacy of the attribute matrix. Multiple regression analyses—with item difficulties as the dependent variable and the nine attribute vectors of **Q** as the independent variables—were performed for the two parallel (OE) subsets and for the two stem-equivalent (MC) subsets for Forms I and II. The multiple R^2s for the two OE subsets of Form I and II were .94 and .92 (R^2s adjusted for shrinkage were .88 and .84), respectively. The R^2s for the two stem-equivalent subsets for Forms I and II were .68 and .73 (R^2s adjusted were .40 and .50), respectively.

Computer programs. The HYBIL program (Yamamoto, 1991) was used to estimate the discrimination (a) and difficulty (b) parameters of the two-parameter IRT logistic model for the test items (see Table 1). The BUGLIB program (Varadi & Tatsuoka, 1989) was used to derive the ideal-score patterns corresponding to the attribute mastery patterns that constituted the groups into which the students' actual response patterns were classified. As a result, 42 groups (knowledge states) were generated for each of the two test forms. The same program also was used for the classification. Each student was classified three times based on the three subsets of items.

Scoring. Scores on the attributes then were compared across the three subsets. Matches and mismatches were counted across the nine pairs of attributes of the contrasted subsets for each student and classified according to the following primary categories:

E = matched mastery (1,1);
F = mastery/nonmastery (1,0; 0,1); or
G = matched nonmastery (0,0).

Analysis. Phi coefficients (ϕ) were computed for pairs of attributes from the parallel and stem-equivalent subsets. The significance of the differences between the coefficients for the pairs was tested using a t test for dependent samples (Hotelling, 1940). The knowledge state to which the student was classified based on each subset was compared across the three analyses, and the number of matches in each sample for states based on parallel subsets was compared to the

Table 2
The Attributes and Relevant Items

Attribute Number	Attribute Description	Item Number	
		Form 1	Form 2
1	Adding a term to both sides of the equation	8	7
2	Subtracting a term from both sides of the equation	1,4,5,6,7	1,4,5,6
3	Applying arithmetic order of operations	1,5,8	1,7
4	Applying the distributive law	5	5
5	Adding or subtracting variable terms	5,6	5,6
6	Dividing across by the coefficient of x, (resulting in $x = b/a$ when $a < b$)	3,4,5,6,7	3,4,5,6
7	Dividing across by the coefficient of x, (resulting in $x = b/a$ when $a > b$)	2	2
8	Applying symmetry law	3,5,7	3,5
9	Evaluating the equation to determine the simplest solution path	5	5

number of matches based on stem-equivalent subsets.

Results

Bug Analysis

Table 3 provides an example of bug analysis results at the individual level. The table contains an individual student's response vectors on the three subsets of Form I; 1 denotes the correct answer, and the letters denote different bugs. On Subsets 1 and 2 (the two parallel OE subsets), the student had seven matched correct responses and one matched bug. Thus, the percentage of matched correct responses on parallel items for this student was 87.5, and the percent of matched bugs was 12.5. On the stem-equivalent subsets (Subsets 1 and 3), the student had five matched correct responses (62.5%), one matched bug (12.5%), and two one-correct-one-error pairs that account for 25% of the eight paired responses.

Table 4 shows the percentage of matched and nonmatched responses across Subsets 1 and 2 and Subsets 1 and 3 for the total group. On average, 78% of the responses to parallel items in the Form I group yielded a match (A = 70.5% matched correct responses and C = 7.5% matched bugs). For the OE vs. MC contrast, the mean overall match was 64.9% (A = 59% and C = 5.9%). In the Form II group, an average of 77.4% of the responses to parallel items yielded a match (A = 69.3% and C = 8.2%). For the OE vs. MC contrast, the overall match was 70.1% (A = 65.4% and C = 4.6%).

For correct/incorrect scoring, the mean overall match of correct (1,1) and incorrect (0,0) responses to parallel items in the Form I group (A + C + D) was 87.3%, with the rest of the item pairs having one correct and one incorrect answer. Thus, of the incorrect pairs (0,0), an aver-

Table 3
Response Vectors for a Student on the Three Subsets of Form I (1 Denotes the Correct Answer and the Letters Denote Different Bugs)

	Item							
Subtest	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	X
2	1	1	1	1	1	1	1	X
3	1	Y	1	1	Z	1	1	X

Table 4
Mean (M) and SD of Percentage of Matched Responses on Two Sets of Parallel (OE/OE) and Stem-Equivalent (OE/MC) Items [A + C = Matched Responses (1,1; Bug = Bug); C + D = Matched Incorrect Responses in 1/0 Scoring Method (0,0); and A + C + D = Matched Responses in 1/0 Scoring Method (1,1; 0,0)]

Form, Item Type, and Statistic	Category of Response						
	A	B	C	D	A + C	C + D	A + C + D
Form I (N = 117)							
OE/OE							
M	70.5	12.7	7.5	9.3	78.0	16.8	87.3
SD	28.5	14.1	11.2	19.4	24.8	23.7	14.1
OE/MC							
M	59.0	20.5	5.9	14.6	64.9	20.5	79.5
SD	27.5	16.0	10.1	21.7	24.7	25.0	16.0
Form II (N = 114)							
OE/OE							
M	69.3	11.0	8.2	11.5	77.4	19.7	89.0
SD	30.7	12.0	18.4	24.2	26.8	28.0	12.0
OE/MC							
M	65.4	18.2	4.6	11.8	70.1	16.4	81.8
SD	31.2	18.3	10.0	21.0	27.1	26.0	18.3

age of 44.6% consisted of matched bugs [C/(C + D)]. For the OE vs. MC contrast, the mean overall match of correct and incorrect responses was 79.5%, with the rest of the item pairs having one correct and one incorrect answer. Thus, of the incorrect pairs, an average of 28.8% consisted of matched bugs. For the Form II group, the parallel subsets contrast yielded a mean of 89% matched responses. Of the incorrect pairs, an average of 41.6% were matched bugs. For the OE vs. MC contrast, an average of 81.8% of the responses matched, and 28% of the incorrect pairs were matched bugs.

Rule-Space Analysis

Table 5 provides an example of the rule-space analysis at the individual level, based on the responses given by the student whose bug analysis was presented above. Table 5 contains the three vectors of the nine attributes as derived from the responses to the three subsets. A comparison of the two attribute row vectors based on parallel items (Subsets 1 and 2) indicates that they are identical. That is, they reflect the same knowledge state (Knowledge State 2). Of the nine attributes, the student mastered eight; thus, the percentage of matched mastery attributes (1,1) for this student was 88.9, the percentage of matched nonmastery was 11.1, and the percentage of one mastery and one nonmastery was 0.

This student's response patterns to the test items perfectly matched (i.e., had a Mahalanobis distance D^2 of 0.0 from) Knowledge State 2, indicating nonmastery of Attribute 1. The two attribute row vectors based on stem-equivalent items (Subsets 1 and 3) matched on five (55.6%) mastered attributes and on one (11.1%) nonmastered attribute. The remaining 33.3% were classified into the one mastery and one nonmastery category. According to the attribute mastery pattern based on Subset 3, the student's pattern perfectly matched Knowledge State 16, indicating nonmastery of Attributes 1, 4, 7, and 9.

Table 6 shows the mean and standard deviation of percentage of matched and nonmatched responses across the nine pairs of attributes for Forms I and II. For the Form I group, a mean of 83.5% of the attributes based on the parallel subsets (OE/OE) yielded a match (E = 68.7% for mastery and G = 14.8% for nonmastery). The mean overall match for attributes based on the two stem-equivalent (OE/MC) subsets was 73.4% (55.8% mastery and 17.7% nonmastery). The correlation between the mastery scores, which is an index of the reliability of these scores, was $r = .71$ for the two parallel (OE/OE) subsets and $r = .53$ for the two stem-equivalent (OE/MC) subsets. For the Form II group, the average match was 85.1% based on the parallel subsets (69.5% for mastery and 15.6% for nonmastery). The average match for attributes based on the OE/MC subsets was 78% (65.5% for mastery and 12.5% for nonmastery). The reliability correlations for the attributes based on the OE/OE and OE/MC subsets were $r = .76$ and .60, respectively.

Table 7 presents the phi coefficients (ϕ) between pairs of attributes from the two OE subsets and from the OE vs. MC subsets. Eight of the nine comparisons for Form I and five of the nine for

Table 5
Response Vectors on the Nine Attributes From Responses to the Three Subsets of Form I for a Student and Mahalanobis Distance (D^2) in the Rule Space Between the Student's Point and the Centroid of the Closest Knowledge State

	Attribute									Knowledge	
Subset	1	2	3	4	5	6	7	8	9	State	D^2
1	0	1	1	1	1	1	1	1	1	2	0.0
2	0	1	1	1	1	1	1	1	1	2	0.0
3	0	1	1	0	1	1	0	1	0	16	0.0

Table 6
Mean (M) and SD of Percentage of Matched Attributes Based on Responses to Two Sets of Parallel (OE/OE) and Stem-Equivalent (OE/MC) Items [E + G = Matched Mastery + Matched Nonmastery (1,1;0,0)]

Form, Item Type, and Statistic	Category of Response			
	E	F	G	E + G
Form I (N = 117)				
OE/OE				
M	68.7	16.5	14.8	83.5
SD	27.9	17.4	22.8	17.4
OE/MC				
M	55.8	26.6	17.7	73.4
SD	24.9	16.2	21.1	16.2
Form II (N = 114)				
OE/OE				
M	69.5	14.9	15.6	85.1
SD	30.7	19.0	25.5	19.0
OE/MC				
M	65.5	22.0	12.5	78.0
SD	30.8	21.3	21.9	21.3

Form II yielded significant differences—all in one direction—indicating a higher match on mastery and nonmastery for attributes based on parallel OE items than for those based on stem-equivalent OE vs. MC items.

Finally, the rule-space classification results based on the different subsets were compared. In Form I, 12.8% of the students were classified into the same knowledge-state group based on their responses to the two stem-equivalent (OE vs. MC) subsets; 38.5% were classified into the same knowledge state group based on their responses to the two parallel OE subsets. The corresponding results for Form II were 28.9% for the stem-equivalent and 42.1% for the parallel subsets.

Discussion

The results of the bug and rule-space analyses yielded similar results with respect to the format effect. Both analyses indicated a closer similarity between the two parallel OE subsets than between the stem-equivalent OE and MC subsets. This was the case for the attribute accounts of item difficulties (as indicated by the regression results), for the knowledge state classification results, for the single attribute mastery level comparisons, and for the response categories at the item level. It also held true for the item difficulties and the subset reliabilities. On the average, the MC subsets tended to be more difficult and had lower internal consistency reliabilities than the two OE subsets.

Most MC items were more difficult than their OE counterparts, which could have resulted from the inclusion of common "bugs" as distractors. Some students correctly answered the OE version of an item but incorrectly answered the MC version—perhaps because they did not bother to engage themselves in the entire "tedious" calculation, but instead tried to take a shortcut by consulting the information in the distractors and hence fell into the trap of selecting an incorrect answer that seemed reasonable. This seems more likely than guessing as an explanation of the strategy used by students in approaching the MC items in this study, because guessing would have resulted in a higher rate of correct responses for the MC than for the OE items, which was not the case. The fact that several studies have found MC items to be easier than their OE counterparts (e.g., Martinez, 1991; Martinez & Katz, 1992; Oosterhof & Coats, 1984) may be explained by the effect of guessing. Thus, the inconclusive results regarding the effect of response format on item difficulty, as noted by Traub & MacRury (1990), may lie in the type of distractors, as well as in the type of task being tested. However, it should be noted that both explanations for students' strategic behavior in answering MC items would affect the reliability of the MC items—rendering it lower than that of a stem-equivalent OE set of items, a fact that was evident in the current study as well as in other studies (e.g., Ackerman & Smith, 1988; Birenbaum & Tatsuoka, 1987; Oosterhof & Coats, 1984; Zimmerman, Williams, & Symons, 1984).

The issue of format effect on item difficulty and reliability is of little interest if the two formats are measuring different abilities or pose different information-processing requirements. Much research has been devoted to the issue of

Table 7
Phi Coefficients (ϕ) for Attribute Pairs Based on Sets of Parallel and Stem-Equivalent Items for Forms I and II

Parallel Formats Attribute Pairs (OE)		Stem-Equivalent Attribute Pairs (OE/MC)		Significance of Difference
Pair	ϕ	Pair	ϕ	
Form I: $N = 117$, $n = 8$				
1	.51**	1	.22**	2.89**
2	.68**	2	.54**	2.16*
3	.50**	3	.30**	2.09*
4	.40**	4	.15**	2.31*
5	.47**	5	.44**	.36
6	.75**	6	.35**	6.33**
7	.54**	7	.35*	2.17*
6	.55**	8	.22**	3.28**
9	.40**	9	.15**	2.31*
Form II: $N = 114$, $n = 7$				
1	.42**	1	.25**	1.56
2	.52**	2	.45**	1.08
3	.67**	3	.41**	4.03**
4	.56**	4	.42**	1.73
5	.66**	5	.43**	3.19**
6	.68**	6	.41**	3.72**
7	.51**	7	.23*	2.95**
8	.68**	8	.26**	4.78**
9	.56**	9	.42**	1.66

*$p < .05$; **$p < .01$ (2-tailed)

trait-equivalence between OE and MC formats. Based on a comprehensive literature review, Traub & MacRury (1990) concluded that different abilities seem to be demanded by the two formats, but the nature of the difference is not well understood. Evidence from protocol analyses of students solving OE and MC items could provide more insight into the nature of the information-processing requirements of the two formats. Martinez & Katz (1992) incorporated protocol analyses in their study of constructed figural response and MC items in architecture assessment. They concluded that format differences in item statistics occurred when different processes were used to solve the two item types.

The nature of the different processes involved in solving MC and OE items needs to be further investigated. Other areas requiring further study include the extent to which the distinction between two formats is generalizable to other domains, what characterizes the items that require different processes, and what characterizes the students that use the different processes. However, the present study indicated that the two types of formats may lead, in some cases, to different diagnoses regarding the student's mal-rules or nonmastered attributes. The fact that the diagnostic inferences from the OE items tended to be more similar across identical formats with parallel items than across different formats with stem-identical items suggests that bug instability is not the cause of the differences in the diagnostic inferences resulting from the two formats. Using Messick's (1989) definition of construct validity as "an integration of any evidence that bears on the interpretation or meaning of the test scores" (p. 17), the results of the present study suggest that compared to the MC format, the OE format provides a more valid measure for the purpose of diagnostic assessment.

References

Ackerman, T. A., & Smith, P. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12,* 117–128.

Bennett, R. E. (1991). *On the meanings of constructed response* (Research Rep. No. RR-91-63). Princeton NJ: Educational Testing Service.

Bennett, R. E., Rock, D. A., Braun, H., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement, 14,* 151–162.

Birenbaum, M., & Shaw, D. J. (1985). Task specification chart: A key to a better understanding of test results. *Journal of Educational Measurement, 22,* 219–230.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response format—it does make a difference. *Applied Psychological Measurement, 11,* 385–395.

Bridgeman, B. (1991). Essays and multiple-choice tests as predictors of college freshman GPA. *Research in Higher Education, 32,* 319–332.

Bridgeman, B., & Lewis, C. (1991). *Sex differences in the relationship of advanced placement essay and multiple choice scores to grades in college courses* (Research Rep. No. RR-91-48). Princeton NJ: Educational Testing Service.

Brown, J. S., & Burton, R. B. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2,* 155–192.

Gutvirtz, Y. (1989). *Effects of sex, test anxiety and item format on performance on a diagnostic test in mathematics.* Unpublished M.A. thesis, School of Education, Tel-Aviv University (in Hebrew).

Hotelling, H. (1940). The selection of variables for use in prediction with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics, 11,* 271–283.

Martinez, M. (1991). A comparison of multiple-choice and restricted figural response items. *Journal of Educational Measurement, 28,* 131–145.

Martinez, M. E., & Katz, I. R. (1992). *Cognitive processing requirements of constructed figural response and multiple-choice items in architecture assessment* (Research Rep. No. RR-92-5). Princeton NJ: Educational Testing Service.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 13–103). New York: Macmillan.

Oosterhof, A. C., & Coats, P. K. (1984). Comparison of difficulties and reliabilities of quantitative word problems in completion and multiple-choice item formats. *Applied Psychological Measurement, 8,* 287–294.

Payne, S. J., & Squibb, H. R. (1990). Algebra mal-rules and cognitive accounts of error. *Cognitive Science, 14,* 445–481.

Sleeman, D., Kelly, A. E., Martinak, R., Ward, R. D., & Moore, J. L. (1989). Studies of diagnosis and remediation with high school algebra students. *Cognitive Science, 13,* 551–568.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 34–38.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95–110.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 50,* 55–73.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive analysis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. C. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 543–488). Hillsdale NJ: Erlbaum.

Tatsuoka, K. K. (1991). *Boolean algebra applied to determination of universal set of knowledge states* (Research Rep. No. ONR-1). Princeton NJ: Educational Testing Service.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7,* 81–96.

Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and pattern classification. *Psychometrika, 52,* 193–206.

Traub, R. E., & MacRury, K. (1990). Antwort-auswahl vs freie-antwort aufgaben bei lernerfolstests [Multiple choice vs. free-response in the testing of scholastic achievement]. In K. Ingenkamp & R. S. Jäger (Eds.), *Tests und trends 8: Jahrbuch der pädagogischen diagnostik* (pp. 128–159). Weinheim, Germany: Beltz Verlag.

Van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement, 14,* 1–12.

Varadi, F., & Tatsuoka, K. K. (1989). *BUGLIB* [Unpublished computer program]. Trenton NJ.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement, 6,* 1–11.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine scorable forms of a test. *Journal of Educational Measurement, 17,* 11–29.

Yamamoto, K. (1991). *HYBIL: Hybrid model of IRT and latent classes* [Computer program]. Princeton NJ: Educational Testing Service.

Zimmerman, D. W., Williams, R. H., & Symons, D. L. (1984). Empirical estimates of the comparative reliability of matching tests and multiple-choice tests. *Journal of Experimental Education, 52,* 179–182.

Acknowledgments

The authors thank Maurice M. Tatsuoka for his valuable comments on an earlier draft of the manuscript.

Author's Address

Send requests for reprints or further information to Menucha Birenbaum, Tel-Aviv University, School of Education, Ramat-Aviv 69978 Israel or to E-Mail BIREN@CCSG.TAU.AC.IL.