Open-Ended Versus Multiple-Choice Response Formats—It Does Make a Difference for Diagnostic Purposes

Menucha Birenbaum
Tel-Aviv University

Kikumi K. Tatsuoka
University of Illinois

The purpose of the present study was to examine the effect of response format—open-ended (OE) versus multiple-choice (MC)—on the diagnosis of examinee misconceptions in a procedural task. A test in fraction addition arithmetic was administered to 285 eighth-grade students, 148 of whom responded to the OE version of the test and 137 to the MC version. The two datasets were compared with respect to the underlying structure of the test, the number of different error types, and the diagnosed sources of misconception (bugs) reflected in the response patterns. The overall results indicated considerable differences between the two formats, with more favorable results for the OE format.

The effect of item format on examinee responses has been studied extensively in the past decade. The equivalence of open-ended (OE) items (also known as free-response or recall items) and multiple-choice (MC) items (also known as recognition items) has been addressed by psychometricians and cognitive psychologists. From an information-processing point of view, different models for the two response formats have been suggested (e.g., Bender, 1980). The commonly held view suggests that recall items require examinees to both search for and retrieve information, whereas recognition items require them only to discriminate among the presented information.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 4, December 1987, pp. 385–395

0146-6216/87/040385-11$1.80

Comparisons between the two formats have used various criteria, such as success rate (e.g., Estes & DaPolito, 1967; Heim & Watts, 1967; Loftus & Loftus, 1976; White & Carcelli, 1982), item difficulty (e.g., Cook, 1955; Merwin & Womer, 1969), the traits measured by the test (e.g., Traub & Fisher, 1977; Ward, 1982; Ward, Frederiksen, & Carlson, 1980), retention rate (e.g., Duchastel & Nungester, 1982; Kumar, Rabinsky, & Pandey, 1979), and examinees' strategies for preparing for the test (e.g., Freund, Brelsford, & Atkinson, 1969; Kumar et al., 1979; Loftus, 1971; Tversky, 1973). The results of these studies, however, seem quite inconclusive. A few indicate an advantage for the open-ended format, while others show no difference between the two formats.

Most of these studies used tests of reading comprehension or vocabulary, domains in which the cognitive mechanisms underlying recall versus recognition are applicable. One question that arises, however, is whether the same distinction holds for MC and OE items in procedural tasks such as arithmetic operations. Moreover, none of the studies compared OE to MC items with respect to information from incorrect responses.

Current research into the diagnosis of examinee misconceptions has focused on procedural tasks (see Birenbaum & Shaw, 1985; Birenbaum & Tatsuoka, 1982, 1983; Brown & Burton, 1978; Burton, 1981; Marshall, 1980; Matz, 1980; Tatsuoka, 1983, 1985, 1986a; Van Lehn, 1981). A recommended procedure for designing a MC test is to

consider examinee responses to an OE test, and then construct distractors based on the frequencies of the incorrect responses to the OE test. However, the equivalence of the two formats for diagnostic purposes has not yet been examined. This issue is addressed by the present study, which evaluated the effect of the response format (MC vs. OE) on the rules of operation underlying examinees' response patterns in fraction-addition arithmetic items.

Method

The Sample

The sample consisted of 285 eighth-grade students from a junior high school in a Midwestern town in the U.S. All students completed the classroom instruction on fraction addition prior to taking the test. An OE test in fraction addition was administered to 148 students. The remaining 137 students were administered the MC version of the test. Both groups consisted of students attending the mathematical laboratory at the school in two consecutive years. (The two samples are representative of the eighth-grade population in that school because they consist of students with various degrees of competence in mathematics. Scheduling conflicts, interest in using computer-aided instruction, or poor mathematical performance are the main reasons for attending the laboratory classes.)

The Test

The OE version of the test consisted of 38 items with 2 parallel sets of 19 items each. The MC version consisted of the same items as the OE version with 5 distractors each. The distractors were constructed on the basis of a frequency count of errors committed by students on the OE version in a pilot study. Both versions of the test were administered in paper-and-pencil form, with items in the same order. The test design was based on an extensive task analysis of fraction addition operations. The items were selected to represent typical errors identified in previous studies (Tatsuoka, 1984a). (For more details on the test design, see Birenbaum & Shaw, 1985; Klein, Birenbaum, Standiford, & Tatsuoka, 1981.)

Analysis

Three types of analyses were used in order to examine the effect of the response format. The first focused on the underlying structure of the two test forms. Using correct/incorrect item scores, Guttman-Lingoes smallest space analysis (SSA-I; Lingoes, 1972) was employed. This nonmetric multidimensional scaling procedure maps the items into points in Euclidean space. Correlations are employed as measures of proximity between the items, in order to determine the corresponding interpoint distances. The level of measurement is ordinal; therefore, a monotone transformation is applied to the correlations in order to maximize the goodness of fit of the solution to the data (as measured by the coefficient of alienation) in a minimal number of dimensions. The solutions are usually very parsimonious, and an adequate solution is frequently obtained in two or three dimensions for quite complex sets of data (Schlesinger & Guttman, 1969). The interpretation of the results depends upon the configuration of the points.

The second analysis compared the two response formats with respect to the types of errors committed by the students. The computer program SPBUG (Baillie & Tatsuoka, 1983) was used to generate responses to each item. The items were first expressed by an algebraic relation of 6 variables, ($a\ b/c\ +\ d\ e/f$), where a and d represent whole numbers and b/c and e/f are fractions; then, based on a logical error analysis (Klein et al., 1981), different possible combinations of erroneous algebraic derivations were programmed. SPBUG then diagnosed the sources of error by matching students' responses with the responses generated by the program.

A previous analysis of the OE version of the test, using SPBUG, identified 70 different error types, which accounted for 80% of the free responses (Tatsuoka, 1984a). Because the MC test included only 20 different error types, the OE responses were coded twice, once according to all the identifiable

error types and once according to those represented in the MC test. Thus, two datasets for OE responses were created. The first dataset will be referred to henceforth as OE, whereas the second will be referred to as OE/MC (i.e., the OE dataset coded according to the distractors of the MC test). The OE and OE/MC datasets were compared to the MC dataset with respect to the mean number of different error types.

The third analysis focused on the entire response pattern in an attempt to diagnose the students' sources of misconception (''bugs'') with respect to fraction addition operations. (As used here, bug denotes a response pattern that includes a set of error types which result from ill-composed rules of operation. For examples of bugs in fraction operations see Birenbaum & Shaw, 1985; Tatsuoka, 1984a, 1986a.) The ''rule space'' method developed by Tatsuoka (Tatsuoka, 1983, 1984b, 1985) was employed for this analysis.

The rule space is a method that classifies response patterns into probability ellipses representing erroneous rules of operation, or bugs (Tatsuoka & Tatsuoka, 1987). All the response patterns resulting from both the correct and the erroneous rules are mapped into a two-dimensional space, where one dimension represents the level of the latent ability being measured by the test (θ in IRT terms), and the other is an index of the atypicality of the response pattern (ζ; Tatsuoka, 1984b). These two parameters were calculated for the dataset at hand, using the two-parameter logistic model. The values of θ and ζ for a simulated response pattern for each bug, selected from a ''bug library,'' are mapped into this space creating the center of each probability ellipse. The actual response patterns are then classified into these ellipses. (See Tatsuoka & Tatsuoka, 1987, for details about the classification procedure.)

The bug library employed in the current study was constructed using two contrasting methods. One method was based on combinations of different item types: addition of simple fractions or mixed numbers with like or unlike denominators. The resulting 16 item-type combinations were used to generate the bugs. An item was assigned a value of 1 if it belonged to a certain item type, and a value of 0 otherwise. It should be noted that a few of the response patterns generated by this method are less likely to occur, due to differences in the number of procedural steps involved in solving the various item types.

The other method for generating bugs was based on a rational task analysis (Klein et al., 1981; Tatsuoka, 1986b; Tatsuoka & Chevalaz, 1984). Task components (attributes) underlying every item were determined, and bugs were generated by assigning a value of 1 to all items sharing the same attribute or combination of attributes which were the target for the diagnosis, and assigning a value of 0 to the remaining items. As a result, a total of 35 ellipses were used for classifying students' response patterns. (See Tatsuoka, 1986a, for detailed information about the bugs.)

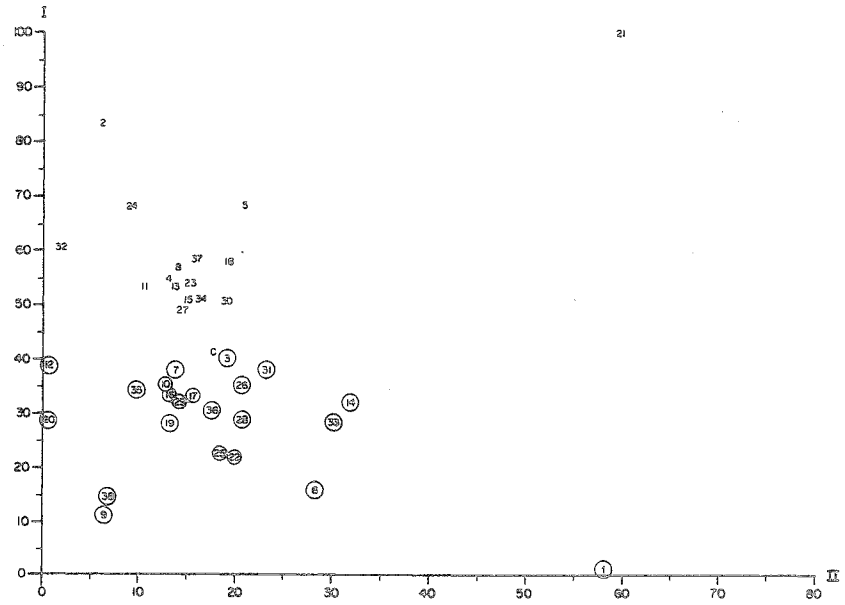
Due to the generating procedure of the first method, a few ellipses were difficult to interpret (e.g., a category of response patterns with correct answers to mixed numbers with unlike denominators, and incorrect answers for the remaining items). Consequently, the 35 ellipses were classified into two categories: those which are easy to rationalize and those that are difficult to rationalize. The OE and MC datasets were compared with respect to these two categories using a χ^2 significance test. Finally, the shortest Mahalanobis generalized distance (D^2) between the point corresponding to the student response pattern in the rule space and the centroid of the bug probability ellipse (Tatsuoka & Tatsuoka, 1987) was computed for each student. The D^2 means for the OE and the MC groups were compared, as were the θ and ζ parameters, using t-tests. Programs from SPSS (Nie, Hull, Jenkins, Steinbrenner, & Brent, 1975) were used for the statistical analyses.

Results

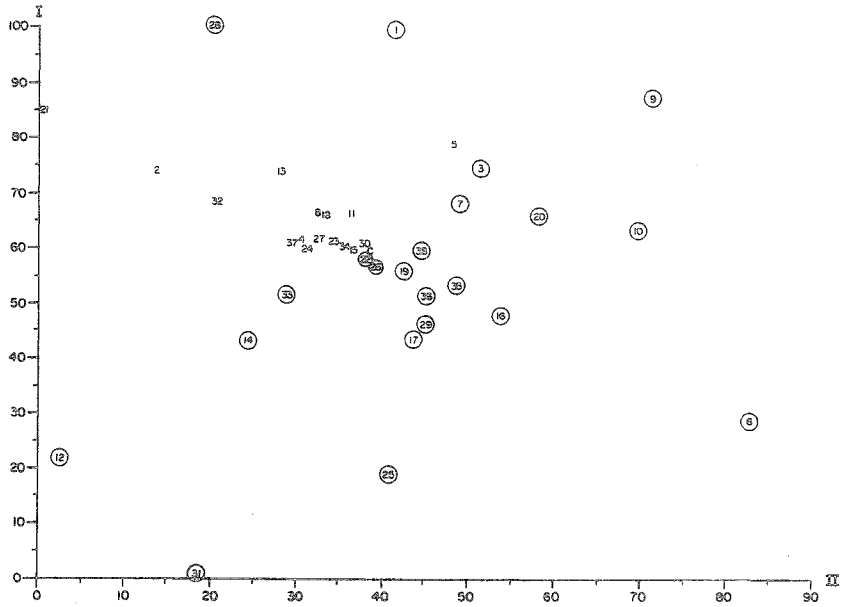
The underlying structure of the test in the OE and MC datasets was examined using smallest space analysis. The two-dimensional solution seemed adequate for both datasets, yielding coefficients of

Figure 1
Two-Dimensional Results From Smallest Space Analysis
(Items With Like Denominators Are Circled)

(a) OE Fraction Addition Test (N = 148)



(b) MC Fraction Addition Test (N = 137)



alienation of .0021 and .0025 for the OE and the MC datasets, respectively. As can be seen in Figure 1a, the OE dataset yielded two distinct clusters of items, the upper one consisting of items with unlike denominators and the lower one consisting of items with like denominators. However, examination of Figure 1b, which represents the MC dataset, shows a more diffused scatter of points in the two-dimensional space with no clear distinction between different types of items.

Table 1 presents the intercorrelation matrices for the two datasets. A factor-analytic examination of those matrices, using the principal factor method, yielded two clear factors in the OE dataset where all items with like denominators loaded on one factor and all items with unlike denominators loaded on the second factor. The factor solution for the MC dataset yielded a less clear distinction; 12 of the items failed to load as expected.

The two datasets were compared with respect to basic test characteristics, including reliability coefficients, total scores, number of omitted items, and number of identified error types. Those results are summarized in Table 2. The α reliability coefficients (Cronbach, 1951) for the two test forms were high (.98 for the OE test and .97 for the MC test), indicating that items within each test form are homogeneous. As can be seen in Table 2, the total test score means for OE and MC did not differ significantly. (Neither did the means of the θ estimates, as will be shown later.) No significant difference was detected between the mean numbers of omitted items in the OE and MC datasets.

The number of identified error types was significantly lower in the two OE datasets (OE and OE/MC) than in the MC dataset. The OE/MC dataset yielded a significantly lower mean than the OE dataset. These results are presented in Table 3. It should be noted that some of the errors in the OE dataset remained unidentifiable by SPBUG. (Tatsuoka, 1984a, reported that the SPBUG identification rate for the OE test was 80% of the incorrect responses.) The OE/MC dataset yielded an even lower identification rate because not all the errors that were identified in the OE dataset were incorporated in the distractors of the MC test.

This difference in error identification rate was taken into consideration when the numbers of different error types in the three datasets were compared. As can be seen in Table 3, even after adjusting the two datasets to compensate for the difference in error identification rate, they yielded significantly lower means than the MC for the number of different error types. A statistically significant difference was also detected between the two OE datasets with respect to this variable, with the OE/MC yielding the lower mean.

These results are on the conservative side because the MC dataset included only 21 response types (20 for incorrect responses and 1 for the correct response), whereas the OE dataset included 70 response types (69 for incorrect responses and 1 for the correct response). If this ratio had been considered, the mean number of error types in the MC dataset would have increased to 22.4. (This calculation takes into consideration that 70 error types were found in a dataset where the identification rate was 80%.) The means to be compared following this adjustment would then be 22.4 for MC, 3.6 for OE, and 1.8 for OE/MC. Obviously, the differences between these means give a much more dramatic appearance to the results.

The third type of analysis examined the differences between the OE and the MC datasets from another perspective. Using the rule space technique, bug ellipses were constructed and classified into two categories of "rationally interpretable" and "rationally uninterpretable" bugs. Table 4 presents the cross-tabulation between these two categories and the two test formats (OE and MC). As can be seen in the table, 84.5% of the response patterns in the OE dataset were classified into the category of rationally interpretable bugs, as compared to only 63.5% of the response patterns in the MC dataset. These differences were significant, as can be seen by the values of the χ^2 statistics.

A further analysis compared the response patterns of the two groups with respect to the two coordinates of the rule space, the ability parameter θ and the atypicality-of-response index ζ, as well as to the shortest Mahalanobis generalized distance (D^2) from the centroid of the bug-ellipse to a given

Table 1
Intercorrelations Among the 38 Test Items in the OE Group (Above the Diagonal) and the MC Group (Below the Diagonal)
(Decimal Points are Omitted)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
1		--	22	39	33	27	39	37	30	38	40	30	36	33	47	31	43	38	37	39	38	21	42	29	32	46	39	35	51	36	44	41	17	44	35	39	45	34	35
2	44		--	42	59	45	32	50	61	35	41	56	41	48	41	60	34	41	59	43	38	37	36	54	50	33	37	60	37	37	60	44	48	48	38	64	48	37	51	34
3	48	44	--	61	54	53	73	60	48	74	65	61	62	58	61	64	65	58	66	54	33	65	57	41	51	65	60	62	67	59	53	49	61	60	56	65	49	50
4	45	53	48	--	66	43	59	82	44	61	82	54	78	52	81	59	56	79	50	50	34	45	82	64	46	60	87	53	57	75	57	71	52	79	60	54	72	40
5	36	30	52	55	--	44	46	71	30	51	57	46	71	41	64	45	45	69	39	43	32	42	64	52	42	49	63	49	45	64	41	54	47	59	43	48	53	33
6	33	21	29	36	37	--	58	41	49	54	41	58	40	49	48	57	57	47	57	47	22	45	50	37	55	56	50	69	59	43	65	40	57	49	58	61	38	49
7	50	43	74	53	53	35	--	57	56	68	60	58	57	58	64	63	67	61	69	54	28	57	58	42	57	67	66	65	67	57	55	47	63	63	61	67	56	55
8	43	48	55	70	61	22	58	--	46	62	78	61	76	50	79	58	54	83	51	51	41	46	81	60	45	53	85	52	55	74	53	75	48	78	55	52	71	38
9	41	43	41	38	41	40	45	36	--	56	54	45	44	46	53	53	53	49	57	45	24	48	46	38	53	43	50	59	52	56	44	41	49	51	50	52	39	55
10	39	31	52	41	35	26	58	46	34	--	61	68	59	61	63	77	76	58	75	58	26	63	60	45	56	61	64	60	76	60	61	48	59	59	65	73	50	57
11	51	60	60	75	51	36	62	76	47	47	--	55	71	53	82	60	56	70	56	53	36	48	75	58	50	50	80	57	60	77	50	65	47	78	52	51	66	43
12	31	31	25	47	27	28	32	39	17	15	32	--	58	51	54	70	61	54	62	50	28	56	56	42	51	55	61	53	62	55	46	53	61	55	46	53	46	45
13	44	48	44	62	49	24	54	72	37	39	68	26	--	52	78	60	54	73	50	53	35	51	79	51	55	55	81	57	55	77	55	67	52	76	55	58	66	40
14	37	42	46	46	40	30	52	59	39	42	49	44	51	--	54	59	69	48	59	58	20	53	53	46	50	58	59	58	66	58	54	39	65	53	57	60	43	51
15	46	58	65	68	59	29	67	81	45	46	77	43	67	66	--	64	63	76	57	52	40	53	85	71	57	60	86	61	64	84	62	71	57	88	62	62	78	48
16	40	35	52	52	44	33	51	54	35	39	43	37	52	56	60	--	76	53	67	67	25	64	61	47	59	60	68	69	53	61	68	51	59	60	70	73	54	59
17	30	34	48	52	42	23	57	58	28	65	54	33	51	55	60	50	--	53	74	63	24	65	58	47	58	69	64	67	80	58	66	50	69	62	69	83	55	67
18	46	52	55	64	57	34	53	75	39	43	70	38	64	51	75	52	50	--	52	43	44	44	75	58	43	54	80	53	51	70	51	66	54	81	54	51	70	37
19	36	44	54	54	44	34	61	59	40	57	60	31	49	53	68	49	65	60	--	55	27	69	54	44	61	60	59	66	71	56	57	49	65	62	62	68	49	65
20	45	40	36	54	47	38	54	45	52	38	51	28	40	36	46	46	45	49	55	--	18	51	51	45	54	57	60	65	62	50	55	48	55	54	56	62	48	50
21	31	61	35	33	32	14	39	49	35	24	45	28	46	43	50	31	28	45	43	32	--	25	42	38	31	26	38	28	26	40	32	32	24	41	20	26	39	17
22	46	48	58	55	51	31	67	67	42	56	61	45	57	56	75	54	55	63	68	48	45	--	46	41	62	68	57	62	70	57	58	37	61	51	58	67	46	66
23	50	56	53	80	60	34	60	75	42	42	77	42	70	54	82	53	59	69	67	58	44	69	--	68	51	59	85	52	62	82	62	75	53	81	59	56	73	42
24	41	51	45	69	56	38	45	67	40	41	66	30	59	47	64	44	44	63	55	54	36	61	77	--	44	51	64	45	47	66	52	58	37	66	45	44	64	41
25	35	28	22	40	33	45	38	39	23	20	40	41	38	38	45	38	36	35	40	40	25	42	53	46	--	63	54	73	66	58	61	41	50	53	65	69	56	59
26	44	47	56	62	54	35	64	68	48	56	65	43	54	62	73	53	54	64	68	54	42	79	72	69	49	--	67	62	71	56	69	55	58	59	68	74	58	63
27	42	48	54	69	64	31	53	79	34	42	75	42	68	49	77	53	53	73	61	51	47	70	83	75	47	71	--	61	65	79	67	73	59	88	66	63	82	47
28	42	40	19	38	41	28	34	42	35	30	41	24	43	33	35	32	30	47	38	48	39	35	43	38	28	31	43	--	67	60	61	47	66	60	63	70	47	58
29	34	45	40	49	38	30	48	54	32	61	55	22	38	47	60	41	62	49	61	46	31	57	56	54	38	59	53	28	--	56	71	50	63	61	72	77	55	70
30	48	59	54	72	64	40	56	73	47	52	78	38	59	53	77	56	56	73	67	60	38	70	80	72	44	77	81	47	65	--	61	64	53	76	59	59	65	49
31	30	24	18	27	17	32	14	23	28	14	30	47	16	40	32	25	33	30	30	27	10	32	31	28	36	34	28	24	28	40	--	53	48	61	79	74	62	57
32	47	60	41	63	45	30	44	53	23	37	56	30	57	34	54	44	36	59	51	45	43	49	65	61	39	52	62	52	44	65	31	--	39	73	55	47	60	42
33	46	56	49	54	45	39	54	48	35	40	58	40	45	48	62	45	45	54	58	43	36	63	60	60	41	59	60	41	54	63	47	61	--	56	57	63	45	54
34	51	58	51	72	59	35	57	77	37	44	78	36	69	50	77	54	54	73	61	57	40	70	83	78	46	72	84	42	63	87	34	68	69	--	64	61	77	49
35	49	41	46	56	47	42	55	54	42	54	59	33	44	42	57	58	55	49	53	56	25	57	63	60	51	68	59	34	66	71	33	50	60	72	--	78	58	64
36	40	41	55	52	45	39	56	55	39	62	55	35	42	52	60	60	72	61	66	50	28	64	57	53	39	69	56	39	68	71	41	45	59	65	75	--	52	70
37	49	57	52	58	56	30	49	65	28	42	65	33	59	43	66	51	46	74	54	47	42	61	63	63	36	61	74	45	51	74	33	77	66	80	56	61	--	42
38	45	47	57	54	47	38	59	57	37	60	57	28	52	45	61	61	54	62	67	50	32	75	58	57	40	69	63	44	64	72	27	58	65	72	67	80	70	--

Table 2
Means, Standard Deviations, and t-Values for Number of Correct Answers, Number of Identified Bugs, and Number of Omissions in the OE and MC Datasets

Variable and Group	N	M	SD
Total Number Correct			
1. OE	148	17.73	13.98
2. MC	137	16.20	13.07
Number of Identified Bugs			
3. OE	148	13.01	12.39
4. MC	137	20.46	13.33
5. OE/MC	148	11.89	12.02
Omitted			
6. OE	148	2.06	6.07
7. MC	137	1.34	4.01

t-values: independent samples
(1) & (2) t = -.95 (283 df)
(4) & (5) t = 5.71** (283 df)
(3) & (4) t = 4.89** (283 df)
(6) & (7) t = -1.18 (283 df)
t-values: dependent samples
(3) & (5) t = 1.29 (147 df)
**p<.01

point corresponding to a student response pattern. As can be seen in Table 5, the differences between the OE and the MC datasets with respect to θ and ζ were insignificant, whereas D^2 yielded a significantly higher value in the MC dataset than in the OE dataset.

These results indicate that the OE group, although not differing significantly from the MC group with respect to ability to solve fraction addition problems, can be better diagnosed with respect to bugs or sources of misconception underlying the response patterns. In order to illustrate the differences in the classification fit between the OE and the MC groups, a few ellipses were chosen and the two groups were plotted against them. Figures 2a and 2b present these results. As can be seen in the figures, the OE dataset captured more of the specified ellipses than did the MC dataset.

Discussion

The results of this study indicated considerable differences between the two formats, with more favorable results for the OE format. The underlying structure, as examined by smallest space analysis, seemed clearer in the OE dataset, where the configuration of the items in the two-dimensional space clearly indicated two clusters: one of items with like denominators and the other of items with unlike denominators. The item configuration for the MC dataset, on the other hand, seemed quite diffuse, with no distinct separation between the different item types.

The results of the error analysis provided an even clearer distinction between the two response formats. Although the two groups did not differ in the ability to solve fraction addition problems, the MC dataset included a significantly larger number of different error types than the OE dataset. This resulted in a less appropriate overall classification rate in the rule space. These results seem to indicate that students who have not mastered the task tend

Table 3
Means, Standard Deviations, and t-Values for Number of Different Error Codes in Each Dataset

Group	N	M	SD
1. OE	148	2.93	2.12
2. MC	137	5.62	3.17
3. OE/MC	148	1.80	1.50
4. OE adjusted[a]	148	3.67	2.65
5. OE/MC adjusted[a]	148	2.26	1.88

[a]Adjusted by adding 1/4 to number of rules (to adjust for 80% bug identification rate in open-ended data).
t-values: independent samples
(1) & (2) t = 8.36** (235 df)
(3) & (2) t = -12.84** (191 df)
(5) & (2) t = -10.81** (283 df)
(4) & (2) t = -5.67** (283 df)
t-values: dependent samples
(1) & (3) t = -10.28** (147 df)
**p<.001

Table 4
Frequencies (N) and Percentages (%) of Test Format by Bug Categories

Test Format	Bug Category: Uninterpretable	Interpretable	Total
OE			
N	23	125	148
%	15.5	84.5	51.9
MC			
N	50	87	137
%	36.5	63.5	48.1
Total			
N	73	212	285
%	25.6	24.4	100.0

Raw $\chi^2_{1df} = 15.32$; corrected $\chi^2_{1df} = 16.40$. Both were statistically significant at $p < .001$

to be less consistent in applying their rules of operation for solving procedural tasks when faced with a MC format than with an OE one.

It seems that the cognitive process involved in these two response formats (in a procedural task) is quite different. According to Fisher and Lipson (1985), "Humans exhibit a fairly strong tendency to avoid extra mental effort, so as to minimize their information processing load" (p. 65). Although in the OE items students had to compute the answer "from scratch," in the MC test they could retrieve cues from the distractors, thus shortening the process, perhaps with more effort directed toward judging the "correctness" of the answer given in the distractors than toward carrying out the entire "tedious" calculation. However, because the distractors in the MC test were carefully chosen to represent common errors rather than random incorrect answers, the task of selecting the correct answer became more complicated and resulted in a wider variety of error types, a greater portion of which were rationally uninterpretable.

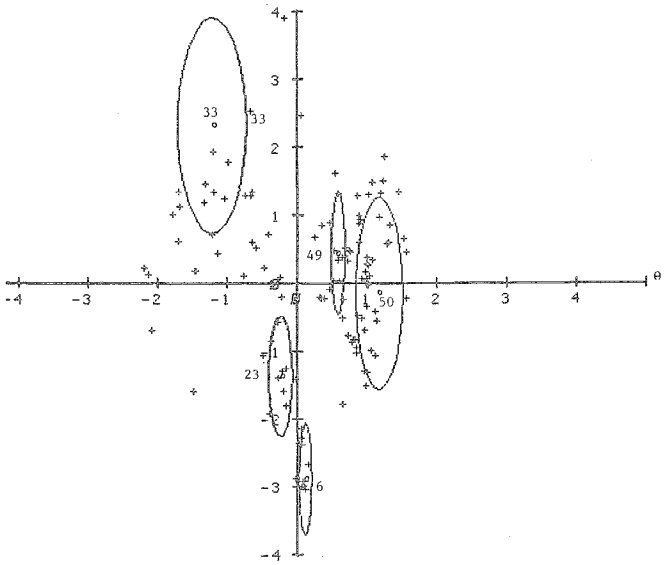
The implications for diagnostic achievement testing in procedural tasks are obvious. MC tests, though considerably easier to score, may not provide the appropriate information for identifying stu-

Table 5
Means, Standard Deviations, and t-Values for θ, ζ, and for the Mahalanobis Generalized Distances (D^2) From the Centroids in the Rule Space for the OE and the MC Datasets

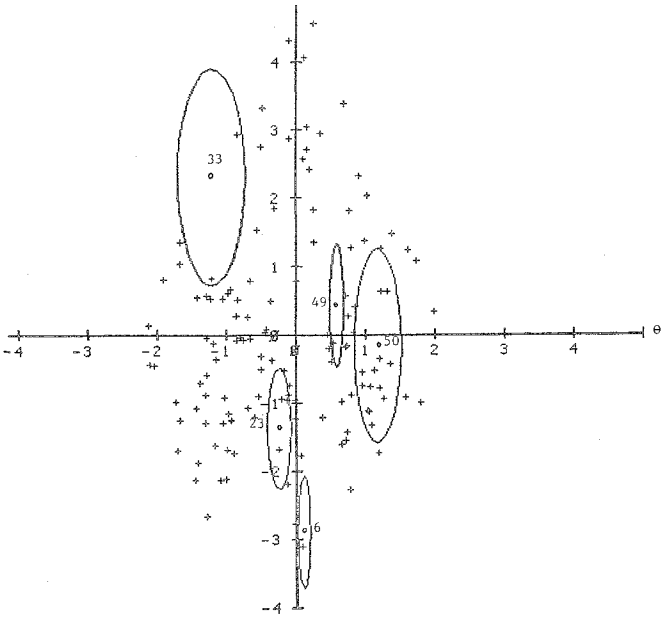
Variable	Group	N	M	SD	df	t	p
θ	OE	148	-.28	1.31	283	-.69	>.05
	MC	137	-.18	1.15			
ζ	OE	148	.18	1.18	258	1.07	>.05
	MC	137	.01	1.50			
D^2	OE	148	.38	.51	245	-2.17	<.05
	MC	137	.54	.71			

Figure 2
Rule Space Plots

(a) OE Dataset



(b) MC Dataset



dents' misconceptions with respect to the given subject matter. The OE format seems more appropriate for this purpose.

References

Baillie, R., & Tatsuoka, K. K. (1983). *SPBUG: A computer program for diagnosing bugs and analyzing responses*. Urbana IL: University of Illinois, Computer-Based Education Research Laboratory.

Bender, T. A. (1980, April). *Processing multiple choice and recall test questions*. Paper presented at the annual meeting of the American Educational Research Association, Boston. (ERIC ED189160)

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. (1982). On the dimensionality of achievement test data. *Journal of Educational Measurement, 19,* 259–266.

Birenbaum, M., & Tatsuoka, K. K. (1983). The effect of scoring based on algorithms underlying the students' response patterns on the dimensionality of achievement test data of the problem solving type. *Journal of Educational Measurement, 20,* 17–26.

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

Burton, R. R. (1981). *Diagnostic bugs in a simple procedural skill*. Palo Alto CA: Xerox Palo Alto Research Center.

Cook, D. L. (1955). An investigation of three aspects of free response and choice type tests at the college level. *Dissertation Abstracts International, 15,* 1351 (University Microfilms No. A55-1791).

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Duchastel, P. C., & Nungester, R. (1982). Testing effects measured with alternate forms. *Journal of Educational Research, 75,* 309–314.

Estes, W. K., & DaPolito, F. J. (1967). Independent variation of information storage and retrieval processes in paired-associate learning. *Journal of Experimental Psychology, 75,* 18–26.

Fisher, K. M., & Lipson, J. H. (1985). Information processing interpretation of errors in college science learning. *Instructional Science, 14,* 49–74.

Freund, R. D., Brelsford, J. W., Jr., & Atkinson, R. C. (1969). Recognition vs. recall: Storage or retrieval differences? *Quarterly Journal of Experimental Psychology, 21,* 214–224.

Heim, A. W., & Watts, K. P. (1967). An experiment on multiple-choice versus open-ended answering in a vocabulary test. *British Journal of Educational Psychology, 37,* 339–346.

Klein, M., Birenbaum, M., Standiford, S. N., & Tatsuoka, K. K. (1981). *On the construction of an error-diagnosing test in fraction arithmetic* (Technical Report 81-6-NIE). Urbana IL: University of Illinois, Computer-Based Education Research Laboratory.

Kumar, V. K., Rabinsky, L., & Pandey, T. N. (1979). Test mode, test instructions, and retention. *Contemporary Educational Psychology, 4,* 211–218.

Lingoes, J. C. (1972). *The Guttman Lingoes nonmetric program series*. Ann Arbor MI: Matesis Press.

Loftus, G. R. (1971). Comparison of recognition and recall in a continuous memory task. *Journal of Experimental Psychology, 91,* 220–226.

Loftus, G. R., & Loftus, E. G. (1976). *Human memory. The processing of information*. Hillsdale NJ: Erlbaum.

Marshall, S. P. (1980). Procedural networks and production systems in adaptive diagnosis. *Instructional Science, 9,* 129–143.

Matz, M. (1980). Towards a computational theory of algebraic competence. *Journal of Mathematical Behavior, 3,* 93–166.

Merwin, J. C., & Womer, F. B. (1969). Evaluation in assessing the progress of education to provide bases of public understanding and public policy. In R. W. Tyler (Ed.), *Educational evaluation: New roles, new means—The sixty-eighth yearbook of the National Society for the Study of Education* (Part II, pp. 305–334). Chicago: University of Chicago Press.

Nie, N. H., Hull, C. H., Jenkins, J. C., Steinbrenner, K. S., & Brent, D. H. (1975). *Statistical package for the social sciences (SPSS)*. New York: McGraw-Hill.

Schlesinger, I. M., & Guttman, L. (1969). Smallest space analysis of intelligence and achievement tests. *Psychological Bulletin, 71,* 95–100.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345–354.

Tatsuoka, K. K. (Ed.) (1984a). *Analysis of errors in fraction addition and subtraction problems* (Final Report for Grant No. NIE-G-81-0002). Urbana IL: University of Illinois, Computer-Based Education Research Laboratory.

Tatsuoka, K. K. (1984b). 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, 10,* 55–73.

Tatsuoka, K. K. (1986a). Diagnosing cognitive errors: Statistical pattern classification based on item response theory. *Behaviormetrika, 19,* 73–86.

Tatsuoka, K. K. (1986b, July). *Toward an integration of item response theory and cognitive diagnosis*. Paper

presented at the ONR Conference on Diagnostic Monitoring of Skill and Knowledge Acquisition, Princeton NJ.

Tatsuoka, K. K., & Chevalaz, M. C. (1984). *A map representation of misconceptions in the rule space: Fraction addition arithmetic* (Research Report 84-2-NIE). Urbana IL: University of Illinois, Computer-Based Education Research Laboratory.

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

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1,* 355–369.

Tversky, B. (1973). Encoding processes in recognition and recall. *Cognitive Psychology, 5,* 275–287.

Van Lehn, K. (1981). *Bugs are not enough: Empirical studies of bugs, impasses and repairs in procedural skills* (Technical Report CIS-11). Palo Alto CA: Xerox, Palo Alto Research Center.

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.

White, K. R., & Carcelli, L. (1982, March). *The effect of item format on students' standardized mathematics achievement test scores*. Paper presented at the annual meeting of the American Educational Research Association, New York. (ERIC ED219425)

Acknowledgments

This research was partially sponsored by the Personnel and Training Research Program, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-82-K-0604, NR 150-495. This research was also partially supported by the National Institute of Education, under Grant No. NIE-G-81-0002. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred. Several of the analyses presented in this report were performed on the PLATO® system. The PLATO system is a development of the University of Illinois and PLATO is a service mark of Control Data Corporation.

Author's Address

Send requests for reprints or further information to Menucha Birenbaum, School of Education, Tel-Aviv University, Ramat-Aviv, Tel-Aviv 69978, Israel.