DOCUMENT RESUME

ED 243 954 TM 840 270

AUTHOR Siskind, Theresa G.; And Others
TITLE Formatting Issues in Multiple Choice Tests.
INSTITUTION Charleston County School District, North Charleston, S.C.
PUB DATE Apr 84
NOTE 32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 24-26, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Decimal Fractions; Geometry; *Grade 7; Junior High Schools; *Multiple Choice Tests; *Performance Factors; *Quantitative Tests; Scores; *Test Format; Test Items

ABSTRACT

The purposes of the present study were to ascertain whether or not two particular formatting characteristics influenced seventh grade student performance on math tests. One research question focused on the effects of decimal positioning in multiple choice options for computational items. In a second question, labeling format was compared with identification format for geometry items. The influences of prior achievement and test composition were also studied. Results indicate that prior achievement had a strong, consistent influence on student performance for both computational and geometry items. Test composition (single versus mixed format) had little influence. Results for formatting are not clearly prescriptive, although additional considerations for geometry formatting are identified. (Author)

ED243954

Research Report

Office of Research and Evaluation
Janet S. Rose, Dir.

Formatting Issues in Multiple Choice Tests

Theresa G. Siskind
William C. Gustin
Pin Pin Tee
Janet S. Rose

Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, April 1984.

TM 820 278

Charleston County School District

Ronald A. McWhirt
Superintendent

Abstract

The purposes of the present study were to ascertain whether or not two particular formatting characteristics influenced seventh grade student performance on math tests. One research question focused on the effects of decimal positioning in multiple choice options for computational items. In a second question, labeling format was compared with identification format for geometry items. The influences of prior achievement and test composition were also studied.

Results indicate that prior achievement had a strong, consistent influence on student performance for both computational and geometry items. Test composition (single versus mixed format) had little influence. Results for formatting are not clearly prescriptive, although additional considerations for geometry formatting are identified.

Formatting Issues in Multiple Choice Tests

Idiosyncracies of multiple choice tests may impede the accurate assessment of student achievement (Millman, Bishop, & Ebel, 1965). To control for the effects of test characteristics on test performance, psychometricians have suggested guidelines for writing items and organizing items within a test (see, for example, Gronlund, 1981; Popham, 1981; Sax, 1974). In addition, some test constructors also attempt to standardize and/or publish their test item format specifications (see, for example, State of Florida, 1982).

The bases for some of these guidelines and specifications have been empirically documented. Organizational factors such as answer marking format (Majors & Michael, 1975; Moss & Cole, 1982), workspace arrangement (Majors & Michael, 1975), sequencing of items (Plake, Ansorge, Parker & Lowry, 1982; Towle & Merrill, 1975) and their interaction with examinee characteristics (Plake, et al., 1983; Towle & Merrill, 1975) have been studied. The effects of item format differences in math items have been examined in terms of specific characteristics such as illustration (Washington & Godfrey, 1974), "set-up" versus computational responses (Forsythe & Spratt, 1980), number of response options (Oosterhof & Coats, 1981), use of "none of the above" as an option (Forsythe & Spratt, 1980; Oosterhof & Coats, 1981), and the use of mathematical ranges as options (Oosterhof & Coats, 1981) as well as more general differences (White & Carcelli, 1982). White and Carcelli contrasted eight commonly used math item formats and concluded, "These results indicate that what a student appears to have mastered in mathematics conputation is substantially influenced by the format of the particular test used" (1982, p. 1).

The two issues addressed in this paper - justification of decimal responses and identification versus labeling - appear to have limited empirical precedents.

Justification of Decimal Responses

Formatting decisions are made daily by test constructors. Scrutiny of nationally standardized tests leads to the conclusion that decimal responses to multiple choice items are not formatted in a uniform way. Decimal options are left-justified in some tests (e.g., The Stanford Test of Academic Skills as in Format B below) and decimal-justified in others (e.g., The Comprehensive Tests of Basic Skills as in Format A below). One purpose of this study was to investigate the effects of decimal formatting on student performance.

Format A	Format B
1.2 x 3.45 = ____	1.2 x 3.45 = ____
a) .935	a) .935
b) 4.14	b) 4.14
c) 9.25	c) 9.25
d) 41.30	d) 41.30

Identification versus Labeling

Not infrequently, test developers use "identification" and "labeling" items to test the same skill domain. For example, an item pool assessing a geometry objective which requires students to identify polygons by name might contain both of the following items:

Item 1 (Labeling):

What is the name of the following figure? 

a) hexagon c) pentagon

b) octagon d) quadrilateral

Item 2 (Identification):

Which figure is a pentagon?

(a) (b) (c) (d)

The second purpose of this study was to compare student performance on items of these types, and explore whether the items actually assess the same domain.

Research Questions

Two mediating variables were identified: previous achievement levels of the sample, and test composition. Test composition refers to the formats of items within a test. A test might include items of a single format, e.g., all items are decimal-justified. Or a test might include items of mixed formats, e.g., some items are decimal-justified and some items are left-justified.

Consideration of the formatting problems and mediating variables led to the following research questions:

1. Are scores on mathematics tests influenced by the decimal positioning (formatting) of responses?

   1.a. Is the effect of decimal formatting constant across achievement levels?

   1.b. Is the effect of decimal formatting influenced by test composition? Would scores of a test with all items formatted similarly differ from scores of a test with mixed item formats?

2. Are scores on mathematics tests influenced by the formatting of items in an identification format versus a labeling format?

   2.a. Is the effect of geometry formatting constant across achievement levels?

2.b. Is the effect of geometry formatting influenced by test composition? That is, will scores on singly formatted tests (identification format only or labeling format only) differ from scores on mixed format tests (identification items mixed with labeling items)?

Method

Procedures

Four forms of a 40-item multiple choice test were constructed and administered to approximately 400 seventh grade students. Computational skills were assessed by twenty-four of the items and geometry concepts by sixteen.

Computational items included six items from each of the following areas: addition of decimals, subtraction of decimals, multiplication of decimals, division of decimals. For each item, two response formats were devised - left justified and decimal justified. Geometry items were also constructed using two different formats - identification and labeling.

The four test forms were constructed in the following manner. On Form 1, all computational items were the decimal-justified versions and all geometry items required labeling. Form 2 consisted of left-justified computational items and geometry identification items. Form 3 was composed by taking the odd-numbered items from Form 1 and even-numbered items from Form 2. Form 4 was composed by taking odd-numbered items from Form 2 and even-numbered items from Form 1.

Geometry items were interspersed among the computational items at regular intervals. Computational items were rotated by content throughout the tests.

Sample

The four forms were randomly assigned to 367 students in twenty-one seventh grade mathematics classes from two suburban middle schools in a large, Southern school district. The two schools were selected on the basis of achievement test scores and student population characteristics. Previous year's achievement test scores indicated that students in both schools represented a wide range of abilities.

Instruments

Estimates of internal consistency reliability were computed for each of the four test forms and for the computation and geometry subtests. KR20 estimates of reliability for the four test forms were 0.81, 0.86, 0.84 and 0.84, respectively. Estimates for the computation subtests ranged from 0.83 to 0.86, and estimates for the geometry subtests ranged from 0.51 to 0.56.

Item analyses were performed separately for computation and geometry subtests. All distractors appeared to perform well.

Mean difficulty estimates for the computation subtests were 0.66, 0.66, 0.68 and 0.64 for the four forms. On Forms 1 and 4, point-biserial correlations between item and total computation subtest scores for two very difficult items fell below +.20. The remaining correlations ranged from +.20 to +.74.

For the geometry subtests, mean item difficulty values were 0.54, 0.56, 0.61 and 0.55. One geometry item displayed point biserial correlations below +.20 on two forms, and one other item fell below +.20 on one form. Remaining values were between +.21 and +.55.

Analysis and Results

Justification of Decimal Responses

A major purpose of this study was to determine whether or not scores on decimal computation items are influenced by formatting of decimal responses. Previous achievement and test composition were recognized and controlled for as possible mediating factors. Conceptually, this design is represented by Figure 1.

Insert Figure 1 about here

Composition of the four forms necessitated an alternate design utilizing subsets of items. On Forms 1 and 3 of the composition subtest, items 1, 5, 6, 7, 11, 12, 13, 17, 18, 19, 23 and 24 were decimal-justified. On Forms 2 and 4, the same items were left-justified. These items formed one subset. On Forms 1 and 4, the other 12 items were decimal-justified, whereas, on Forms 2 and 3, they were left-justified. These items formed a second subset.

In order to address the justification question, scores on decimal-justified items were compared with scores on left-justified items for each subset independently. A two factor analysis of variance, with format as one factor and achievement as the other, was performed. Cell means and standard deviations are reported in Table 1, and results of the analysis of variance are listed in Table 2.

Insert Tables 1 - 2 about here

To isolate the effects of test composition, scores on each subset of items were analyzed across forms. For example, scores on Form 1 (single format) were contrasted with scores on Form 3 (mixed format) for the first subset of items. On both Form 1 and Form 3, the first subtest was comprised of decimal justified items. Similar analyses were performed for other appropriate combinations. Cell means and standard deviations, and ANOVA results are reported in Tables 3 - 6.

Insert Tables 3 - 6 about here

A two factor analysis of variance, with achievement as one factor and test forms as the other, was also performed using computation subtest scores as the dependent variable. Results are summarized by Tables 7 and 8.

Insert Tables 7 and 8 about here

Identification versus Labeling

A second major purpose of this study was to determine whether or not scores on geometry items are influenced by item format - identification or labeling. Previous achievement and test composition were recognized and controlled for as possible mediating factors. Conceptually, this design is represented by Figure 2.

Insert Figure 2 about here

Composition of the four forms necessitated an alternate design utilizing subsets of items. On Form 1 and 3 of the geometry subtest, items 1, 2, 5, 6, 9, 10, 13 and 14 required labeling. On Forms 2 and 4, the same items necessitated identification. These items formed one subset. On Forms 1 and 4, the other eight items were labeling and on Forms 2 and 3, they were identification. These items formed a second subset.

To address the identification versus labeling question, scores on labeling items were compared with scores on identification items for each subset independently. A two factor analysis of variance, with format as one factor and achievement as the other, was performed. Cell means and standard deviations are reported in Table 9, and results of the analysis of variance are listed in Table 10.

Insert Tables 9 and 10 about here

To isolate the effects of test composition, scores on each subset of items were analyzed across forms. For example, scores on Form 1 (single format) were contrasted with scores on Form 3 (mixed format) for the first subset of items. On both Form 1 and Form 3, the first subset was comprised of labeling items. Similar analyses were performed for other appropriate combinations. Cell means and standard deviations and ANOVA results are reported in Tables 11 - 14.

Insert Tables 11 - 14 about here

A two factor analyses of variance, with achievement as one factor and test form as the other, was also performed using geometry subtest scores as the dependent variable. Results are summarized by Tables 15 and 16.

Insert Tables 15 and 16 about here

Discussion

The two stated purposes of the present study were to ascertain whether or not two particular formatting characteristics influenced students performance on math tests. In one instance, the research question was focused on the effects of decimal positioning in the responses of computational items. In the other, labeling format was compared with identification format for geometry items. In addition to the two major questions, the influences of prior achievement and test composition were studied.

Not unexpectedly, previous achievement had a strong, consistent influence on student performance on both geometry and decimal subtests and item subsets. By contrast, test composition appeared to have little effect on student performance.

Geometry

The only exception to test composition effects was for the geometry subtest. Results reported in Table 12 indicate that test composition may have contributed to differential student performance on the first item subset. By viewing these results in conjunction with Table 16, in which test form was a significant main effect, and with item difficulty data, the conclusion may be drawn that Test Form 3 contributed strongly to these findings. A Scheffe post hoc analysis of the Tables 15 and 16 data indicated that Form 3 geometry scores were significantly higher than scores on Forms 1 and 4, and significantly higher than scores on Form 2 for the low and medium achievement groups.

Independent inspection of item difficulty data allowed for a possible explanation. Item difficulty values for each geometry item were compared across the four test forms, averaged across format and derived overall. For the seven items in which the overall difficulty value was .50 or below, five were easier in the labeling format than the identification format. Differences in difficulty across formats were substantial, ranging from 0.06 to 0.24. For the nine items with overall p-values greater than 0.50, seven were easier in the identification format and one was the same. Differences across formats ranged from 0.03 to 0.19. In the composition of Test Form 3, items were randomly selected. Form 3 utilized the easier formats for 12 of the 16 items.

From a different viewpoint, the item analysis data indicated that for assessment of a singular instance of a concept (e.g., pentagon, prism, trapezoid, arc, cylinder, ray, decagon, vertex, rhombus), labeling items tended to be easier than identification items. Differences in difficulty across formats ranged from 0.04 to 0.24 (although one item was equally difficult across formats). For relational and conjunctive concepts (e.g., parallel line segments, obtuse angle, scalene triangle, diagonal, right triangle, on the circle, angle naming), identification items tended to be easier. P-value differences ranged from 0.03 to 0.19. Form 3 was composed of the easier format for 13 of 16 items.

The item analysis data are cited to explain the phenomenon of Form 3 high scores. They may also lead to follow-up studies beyond the purview of the present paper.

Computation

The analyses of the decimal subtest data are somewhat contradictory with regard to format. Table 2 reports no significant main effect for format or for the format x achievement interaction for item subset 1. Yet for item subset 2 the interaction effect is signigicant. Scheffe post hoc analyses indicate a significant difference between decimal and left-justification only for the high achievers with mean scores for the left-justified subtest higher than those for the decimal-justified subtest.

Table 8 indicates a significant interaction effect for form x achievement. The Scheffe analysis indicated differential patterns among the achievement groups across the forms. On Form 1 (decimal-justified format), mean scores indicated that the low and medium achievement groups did not score significantly different from one another on the decimal subtest but were significantly lower that the higher achievers. The same pattern held for Form 3, a mixed format form. For Form 2 (left-justified format), each achievement group scored significantly different from the others. And for Form 4, a mixed format test, the low achievers scored significantly lower than the medium and high achievers, who did not perform significantly differently from one another.

Item analysis data did not indicate specific format patterns in the data either by form, item difficulty or by content (addition, subtraction, multiplication, division). Based on the results of this study, there does not appear to be a rationale for the choice of any particular mode of decimal formatting.

References

Forsythe, R. A., & Spratt, K. F. (1980). Measuring problem solving ability in mathematics with multiple-choice items: The effect of item format on selected item and test characteristics. Journal of Educational Measurement, 17 (1), 31-43.

Gronlund, N. E. (1981). Measurement and evaluation in teaching (4th ed.). New York: Macmillan.

Majors, G. W., & Michael, J. J. (1975). The relationship of achievement on a teacher-made mathematics test of computational skills to two ways of recording answers and two workspace arrangements. Educational and Psychological Measurement, 35, 1005-1009.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test wiseness. Educational and Psychological Measurement, 25, 707-726.

Moss, P. A., & Cole, N. S. (1982, March). A comparison of answer-marking formats for multiple-choice achievement testing in early elementary grades. Paper presented at the meeting of The National Council on Measurement in Education, New York.

Oosterhof, A. C., & Coats, P. K. (1981). Comparison of difficulties and reliabilities of math completion and multiple-choice item formats. Paper presented at the meeting of the American Educational Research Association, Los Angeles.

Plake, B. S., Ansorge, C. J., Parker, C. S., & Lowry, S. R. (1982). Effects of item arrangement, knowledge arrangement, test anxiety and sex on test performance. Journal of Educational Measurement, 19 (1), 1982, 49-57.

Popham, W. J. (1981). Modern educational measurement. Englewood Cliffs, N. J.: Prentice Hall.

Sax, G. (1974). Principles of educational measurement and evaluation. Belmont, CA: Wadsworth.

State of Florida. (1982). Formatting Specifications for the Florida State Student Assessment Tests. Tallahassee, FL: Florida Department of Education.

Towle, M. J., & Merrill, P. F. (1975). Effects of anxiety type and item difficulty sequencing on mathematics test performance. Journal of Educational Measurement, 12 (4), 241-249.

Washington, W. N., & Godfrey, R. R. (1974). The effectiveness of illustrated items. Journal of Educational Measurement, 11 (2), 121-124.

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

	FORMAT			
	Decimal Justified		Left Justified	
ACHIEVEMENT	Decimal Only	Mixed	Left Only	Mixed
High				
Medium				
Low				

Figure 1. Conceptual Representation of Research Question 1

	FORMAT			
	Labeling		Identification	
ACHIEVEMENT	Labeling Only	Mixed	Identification Only	Mixed
High				
Medium				
Low				

Figure 2. Conceptual Representation of Research Question 2

Table 1
Decimal Justified versus Left Justified
Means and Standard Deviations

ITEM SET 1

	Decimal Justified	Left Justified
Low	6.77	5.39
	2.09	2.51
Medium	7.64	7.67
	2.44	2.42
High	9.29	8.71
	1.97	2.16

ITEM SET 2

	Decimal Justified	Left Justified
Low	5.42	5.88
	2.67	2.42
Medium	7.89	7.59
	2.58	2.39
High	9.11	10.30
	2.67	1.72

Table 2
Decimal Format
Decimal Justified versus Left Justified

ITEM SET 1

Source	Sum of Squares	df	F
Format	13.21	1	2.49
Achievement	337.65	2	31.78*
Format X Achievement	23.59	2	2.22

ITEM SET 2

Source	Sum of Squares	df	F
Format	5.18	1	0.89
Achievement	670.50	2	57.59*
Format X Achievement	39.96	2	3.43**

*$p < .01$

*$p < .05$

Table 3
Decimal Justified
Uniform vs. Mixed
Means and Standard Deviations

ITEM SET 1

	Uniform	Mixed
Low	6.90	6.70
	1.91	2.23
Medium	7.81	7.41
	2.34	2.57
High	8.76	9.80
	2.23	1.49

ITEM SET 2

	Uniform	Mixed
Low	6.00	5.00
	2.71	2.66
Medium	7.60	8.27
	2.31	2.89
High	9.21	9.00
	2.85	2.48

Table 4
Decimal Justified
Uniform vs. Mixed

ITEM SET 1

Source	Sum of Squares	df	F
Composition	0.02	1	0.00
Achievement	157.25	2	15.56*
Composition X Achievement	19.94	2	1.97

ITEM SET 2

Source	Sum of Squares	df	F
Composition	0.00	1	0.00
Achievement	227.46	2	16.54*
Composition X Achievement	16.08	2	1.17

*$p < .01$

Table 5
Left Justified
Uniform vs. Mixed
Means and Standard Deviations

ITEM SET 1

	Uniform	Mixed
Low	6.07	4.71
	2.67	2.64
Medium	7.29	8.09
	2.69	2.03
High	9.23	8.04
	1.91	2.31

ITEM SET 2

	Uniform	Mixed
Low	5.71	6.00
	2.58	2.22
Medium	7.55	7.64
	2.55	2.22
High	10.06	10.53
	1.67	1.76

Table 6
Left Justified
Uniform vs. Mixed

ITEM SET 1

Source	Sum of Squares	df	F
Composition	2.62	1	0.49
Achievement	202.76	2	18.82*
Composition X Achievement	45.82	2	4.25

ITEM SET 2

Source	Sum of Squares	df	F
Composition	0.53	1	0.11
Achievement	486.81	2	49.75*
Composition X Achievement	1.36	2	0.14

$*p < .01$

Table 7
Decimal Justification
Means and Standard Deviations

	FORM			
ACHIEVEMENT	1	2	3	4
Low	12.90	11.79	12.70	9.71
	4.15	4.46	3.93	4.78
Medium	15.41	14.84	15.05	16.36
	4.13	4.88	4.46	4.47
High	17.97	19.29	20.33	17.04
	4.92	3.14	3.00	4.40

Table 8
Decimal Justification

Source	Sum of Squares	df	F
Form	28.81	3	0.52
Achievement	1965.46	2	53.57*
Form X Achievement	284.56	6	2.59**

*$p < .01$

**$p < .05$

Table 9
Labeling versus Identification
Means and Standard Deviations

ITEM SET 1

	Labeling	Identification
Low	2.80	2.46
	1.49	1.37
Medium	3.39	2.98
	1.65	1.38
High	4.78	4.13
	1.55	1.48

ITEM SET 2

	Labeling	Identification
Low	4.42	5.38
	1.35	1.48
Medium	5.32	5.47
	1.42	1.45
High	5.92	6.46
	1.47	0.92

Table 10
Geometry
Labeling vs. Identification

ITEM SET 1

Source	Sum of Squares	df	F
Format	20.41	1	8.94*
Achievement	167.70	2	36.72*
Format X Achievement	1.36	2	0.30

ITEM SET 2

Source	Sum of Squares	df	F
Format	14.43	1	7.73*
Achievement	72.51	2	19.41*
Format X Achievement	8.06	2	2.16

*p < .01

Table 11
Geometry Labeling
Uniform vs. Mixed
Means and Standard Deviations

ITEM SET 1

	Uniform	Mixed
Low	2.40	3.00
	1.07	1.65
Medium	3.17	3.68
	1.57	1.72
High	4.45	5.10
	1.40	1.65

ITEM SET 2

	Uniform	Mixed
Low	4.40	4.43
	0.84	1.65
Medium	5.00	5.75
	1.32	1.45
High	5.93	5.92
	1.46	1.50

Table 12
Labeling Format
Uniform vs. Mixed

ITEM SET 1

Source	Sum of Squares	df	F
Composition	12.67	1	5.05**
Achievement	104.89	2	20.90*
Composition X Achievement	0.20	2	0.04

ITEM SET 2

Source	Sum of Squares	df	F
Composition	5.65	1	2.84
Achievement	40.65	2	10.23*
Composition X Achievement	6.09	2	1.53

*$p < .01$

**$p < .05$

Table 13
Geometry Identification
Uniform vs. Mixed
Means and Standard Deviations

ITEM SET 1

	Uniform	Mixed
Low	2.43	2.50
	1.22	1.56
Medium	2.92	3.05
	1.46	1.31
High	4.26	3.96
	1.65	1.23

ITEM SET 2

	Uniform	Mixed
Low	5.36	5.40
	1.28	1.64
Medium	5.43	5.52
	1.71	1.11
High	6.42	6.50
	1.03	0.82

Table 14
Identification Format
Uniform vs. Mixed

ITEM SET 1

Source	Sum of Squares	df	F
Composition	0.20	1	0.10
Achievement	66.27	2	16.40*
Composition X Achievement	1.62	2	0.40

ITEM SET 2

Source	Sum of Squares	df	F
Composition	0.19	1	0.11
Achievement	42.37	2	12.20*
Composition X Achievement	0.02	2	0.00

*p < .01

Table 15
Labeling versus Identification
Means and Standard Deviations

ACHIEVEMENT	FORM 1	FORM 2	FORM 3	FORM 4
Low	6.80	7.79	8.40	6.93
	1.03	1.81	2.72	2.70
Medium	8.17	8.35	9.20	8.80
	2.30	2.31	2.05	2.04
High	10.38	10.68	11.60	9.88
	2.43	2.34	1.90	2.27

Table 16
Labeling versus Identification

Source	Sum of Squares	df	F
Form	70.12	3	4.72**
Achievement	467.12	2	47.14*
Form X Achievement	24.33	6	0.82

*p < .01

**p < .05