DOCUMENT RESUME

ED 219 416 TM 820 427

AUTHOR Siskind, Theresa G.; Anderson, Lorin W.
TITLE The Technical Quality of Test Items Generated Using a Systematic Approach to Item Writing.
PUB DATE Mar 82
NOTE 21p.; Paper presented at the Annual Meeting of the National Council of Measurement in Education (New York, NY, March 20-22, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Algorithms; *Item Analysis; *Multiple Choice Tests; Quality Control; *Response Style (Tests); *Test Construction; *Test Format; Testing Problems; Test Items

ABSTRACT

The study was designed to examine the similarity of response options generated by different item writers using a systematic approach to item writing. The similarity of response options to student responses for the same item stems presented in an open-ended format was also examined. A non-systematic (subject matter expertise) approach and a systematic (algorithmic) approach were the methods of item writing employed. Results indicated that neither approach resulted in very similar response options being generated by writers. Few response options produced by either approach corresponded with the incorrect student responses. The variation in incorrect student responses illustrated that the use of multiple choice items to test basic mathematics operations would not account for all student errors. It is probable that multiple choice test scores would be inflated over open-ended format test scores utilizing the same items. The study highlighted some potential problems in attempting to construct unbiased multiple choice tests. (DWH)

The Technical Quality of Test Items Generated Using a Systematic Approach to Item Writing

Theresa G. Siskind

Charleston (S.C.) County Schools

Lorin W. Anderson

University of South Carolina

Paper presented at the Annual Meeting of the National Council of Measurement in Education, New York, March 19-23, 1982.

The Technical Quality of Test Items Generated Using a Systematic Approach to Item Writing

The intrusion of criterion-referenced testing into the educational scene has led to several interesting and potentially useful developments. One of these developments has resulted from the call for greater specificity of item writing rules and operations if test scores are to become increasingly meaningful. Bormuth (1970) was one of the first to emphasize the need for such increased specificity. According to Bormuth, results obtained from tests made by traditional methods cannot be used as evidence in deciding issues of public policy or instructional theory because "there is no way to refute or confirm the results of a study in which items made by traditional methods were used. The reason for these problems is that items made in the traditional manner are derived through private intuition of the item writer rather than through a set of operations open to public inspection" (Bormuth, 1970, p. 7; emphasis ours).

In response to this identified problem Bormuth devised an approach to item writing which he termed "item transformations." The main purpose of the approach was to increase the similarity of achievement test items written by different item writers to assess student acquisition of the same objectives. Subsequently, a number of educators and testing experts developed other systematic, explicit approaches in order to minimize so-called "item writer bias." Hively et al.'s (1973) item forms, Durnin and Scandura's (1973) algorithmic approach, Popham's (1975) amplified objectives, Berk's (1978) mapping sentences, and Tiemann and Markle's (1978) concept-based testing represent five of the most potentially useful approaches. In a recent review of systematic item writing approaches Roid and Haladyna (1980) endorse the continued and increased use of these approaches while contending that "(m)any of the problems of item-writer bias may be avoided by employing one of several domain-based item-generating approaches" (p. 309).

Despite the magnitude of the problem identified by Bormuth and the endorsement by Roid and Haladyna of systematic item writing approaches as potential solutions to the problem, relatively little research has been conducted to examine the validity of the endorsement. The research conducted to date has employed a traditional index of item difficulty as the dependent variable (Haladyna and Roid, 1978). That is, if items generated by different item writers are associated with similar item difficulties, item-writer bias is said to have been minimized or eliminated.

The use of an index of item difficulty as an indicator of item-writer bias is problematic for at least two reasons. First, if bias is defined in terms of the deviation from a "true" difficulty, then similarity of item difficulties may or may not indicate the elimination of bias. Perhaps both of the item writers employed in one of the Haladyna and Roid studies were equally and systematically biased. If so, little variation in the difficulty of the resulting items, would be a likely result. Second, a comparison of the items themselves (in addition to students' responses to the items) seems necessary if item-writer bias is to be adequately examined and, ultimately, explained.

The issue of item-writer bias for multiple-choice tests includes two facets. First, item stems derived by different item writers from a given content area should be similar as well as representative of the content tested. Second, response options should be replicable across item writers while also mirroring actual student errors. The present study focuses on the latter of these two aspects of item writer bias.

More specifically, the purposes of the study were to examine 1) the similarity of response options generated by different item writers using a systematic approach to item writing, 2) the similarity of response options generated by different item writers who were subject matter specialists and had no knowledge of systematic approaches to item writing, and 3) the similarity of the response options generated by both aforementioned groups of item writers to student responses for the same item stems presented in completion, rather than selection, format.

Samples and Procedures

Three samples were included in the study. Two samples consisted of item writers; a third was composed of sixth grade pupils. The two samples of item writers differed both in terms of their educational background and in the instructions they received relative to the generation of response options.

Sample A consisted of four certified secondary-level teachers holding Masters degrees in mathematics education. These teachers were instructed to examine twenty arithmetic exercises (five problems each for addition, subtraction, multiplication, and division of whole numbers). Based on their knowledge of the way in which sixth grade students learn to work these types of problems, these item writers were asked to generate response options that would correspond to errors such students would be likely to make. Three incorrect response options were to be generated for each exercise resulting in a total of 60 incorrect response options.

Sample B was composed of four undergraduate students enrolled in a test construction course during the next-to-last semester of their senior year. All of these students were majoring in early childhood, elementary, or special education. All had previously completed one course in mathematics education.

One segment of the test construction course was devoted to systematic approaches to item writing. During this segment, one ninety-minute class period was spent on Durnin and Scandura's (1973) algorithmic approach. Briefly, the algorithmic approach is based on the assumption that students use rules to solve problems. By taking account of the rules -- processes and decision points -- that students use to solve problems, a flow chart can be diagrammed for a given type or category of problems.

Four flow charts diagramming procedures for solving addition, subtraction, multiplication and division problems for whole numbers had been constructed by the instructor of the course. (Each flow chart had previously been validated by a teacher review process involving at least three teachers.) The undergraduate students were trained to read the flow charts. In addition, students were told that incorrect response options for math problems could be generated by utilizing the flow charts. In the derivation of options, these students were instructed to make either a process error (an error in performing a step or specified activity) or a decision point error (an error in selecting the appropriate route through the flow chart). The erroneous process or path was to be followed until the algorithm was completed, resulting in an incorrect response. Thus, for each pass through the flow chart, a unique error was to be committed. As in the case of the item writers in Sample A, those in Sample B were instructed to generate 60 response options for the 20 arithmetic exercises.

Sample C consisted of 131 sixth grade pupils enrolled in one of six mathematics classes. Three different teachers, each teaching two classes, instructed the six classes. The same textbook was used in all of the classes. A completion test consisting of only the 20 arithmetic exercises was prepared and distributed to the students. The test directions asked the students to work each problem and write the answer on the test itself. No time limits were imposed for completion of the test.

Findings Concerning Similarity of Response Options Generated by Each Approach

Analyses of the response options generated by the item writers using the algorithmic approach and those having subject matter expertise suggest that neither approach, independently, yielded very similar distracters. Considering the total number of response options generated by both samples of item writers, 71 per cent of those generated by the subject matter experts differed from one another. Similarly, 80 per cent of the responses generated by the item writers employing the algorithmic approach were dissimilar. The number and percentages of different response options developed by each approach are reported for each item and all items in Table 1.

Insert Table 1 about here

Table 1 provides an indication of the variation in generated response options for each approach. However, even if all item writers duplicated each other's distracters, the percentage of different responses would be 25 per cent (indicating each of the four writers had derived the same three options). Table 2 further explicates Table 1 by reporting how many item writers for each approach, separately, derived the same response option.

Insert Table 2 about here

Despite classification by Roid and Haladyna (1980) as one of the more objective item writing approaches, on one plane the algorithmic approach is no less "biased" than more traditional approaches to item writing. In fact, as Tables 1 and 2 indicate, the algorithmic approach tended to result in a greater variety of response options than the non-systematic approach.

When examined from another perspective, however, the algorithmic approach may be well be viewed as relatively "unbiased". If indeed the algorithmic approach does account for rules governing the thought processes, one might expect response options generated by this approach to vary according to the number of processes and decision points present in the correct response algorithm. That is, the more processes and decision points, the greater the chance for student error, and the greater the number of response options that can be generated.

To investigate this possibility, the percentage of dissimilar response options generated by the algorithmically-based item writers was correlated with the number of procedures and decision points present in the correct response algorithm. The resultant correlation was 0.51 ($p < .05$). This finding suggests that the algorithmic approach to response option derivation is sensitive to the complexity of the procedure being tested.

Further analyses resulted in a correlation of 0.55 between the number of dissimilar response options generated by the two samples of item writers. Additionally, the correlation between number of dissimilar response options written by the subject matter specialists and the number of processes and decision points in the correct response algorithms was 0.47. These findings suggest that subject matter specialists may be intuitively following rules in deriving their response options.

Findings Concerning Similarity of Response Options Generated by Both Approaches

In addition to a great deal of "within-approach" variability among response options, a great deal of cross-approach variability was also apparent. Table 3 reports the number of response options generated for each of the 20 exercises by writers for both approaches and the number of writers for each approach deriving the same option(s) common to both approaches.

Insert Table 3 about here

Table 3 illustrates that there is little commonality between response options generated by the two approaches. A closer examination of the actual response options indicates that this finding may be attributable, in part, to inherent differences in the two approaches. In specific, the algorithmic approach discourages individuality in errors and encourages continued commission of an error once it is committed. Thus, for example, in a three-digit subtraction problem with borrowing, once item writers using the algorithmic approach commit a borrowing error, the same error should be committed throughout the problem. Whereas, subject matter specialists may employ their own judgment as to whether or not an error will be continually committed.

Findings Concerning Similarity Between Generated Response Options and Supplied Responses

A comparison of response options generated by writers using both approaches with responses supplied by the sixth-grade students indicates that relatively few of the distracters produced by either sample of item writers correspond with incorrect responses made by the students. Of all incorrect responses made to the 20 exercises, 13 per cent were identical to response options generated by the subject matter specialists and 20 per cent were the same as those produced by the item writers using the algorithmic approach. Table 4 reports the number of incorrect responses made by the students to each item as well as the number and per cent of student-supplied responses corresponding to the response options generated by the two samples of item writers.

Insert Table 4 about here

The algorithmically-derived distracters matched more of the incorrect student responses than options derived by the subject matter experts for 13 of the 20 items. However, neither approach would have resulted in multiple-choice items in which response options matched the majority of incorrect student responses. Furthermore, the analyses summarized in Table 4 utilized all response options generated by item writers (i.e. 10, 11 or 12 response options). Multiple-choice items typically allow only four or five options.

Interestingly, the response options generated by the greatest number of item writers were frequently not the responses which most often matched the correct student responses or the most frequently occurring incorrect student responses. Consider, for example, Item 8 which was answered incorrectly by 38 students. Nineteen (or 50 per cent) of these incorrect responses matched distracters generated by the subject matter specialists (see Table 4). However, the three response options most frequently generated by the subject matter specialists corresponded with only two of the students' incorrect responses. In contrast, response options generated by a single subject matter specialist corresponded with the incorrect responses made by eight students.

This phenomenon is especially problematic with respect to the generation of multiple-choice items. If, in constructing Item 8, one had selected in addition to the correct answer, the three response options generated most frequently by writers, only 95 (73%) of the student responses would have matched the multiple-choice options. If, on the other hand, one had incorporated into Item 8, options less frequently supplied by writers, 110 (84%) student responses would have matched the options.

One major contributor to the lack of congruence between the response options generated by the item writers and the responses supplied by the students is lack of consistency among the incorrect responses of students. Table 5 reports the number of students responding incorrectly to each item and the number and percentage of different responses to each item. A distinction is drawn between the number of students missing the item (N Incorrect) and the number of students who attempted the item and missed it (N Answering). In a test situation, an unanswered question would most likely be marked incorrect. In an analysis of responses, however, questions left unanswered by different students do not necessarily represent the same response or error. Whereas one student may have overlooked the item, another may not have been able to compute the answer.

Insert Table 5 about here.

The analyses summarized by Table 5 show that among students who answer addition, subtraction, multiplication and division problems incorrectly, few arrive at the same incorrect answer. An attempt to analyze the students' incorrect responses across items uncovered only one student whose incorrect responses exhibited a pattern, and for that student, only four of five incorrect responses followed the pattern. Further analyses uncovered little relationship between the number of different incorrect student responses and the number of processes and decision points in a correct response algorithm ($r = 0.18$, $p > .05$). It appears from these preliminary analyses that most student errors in responding to this type of test problem are random; that is, they follow no systematic pattern. Item writers are unlikely to generate random error response options, thus in part, explaining the lack of congruence between student responses and the generated response options. Furthermore, student errors may be compounded, thus partially explaining why student responses were so poorly matched to generated response options, especially algorithmically derived options.

Discussion

Recent item writing technology has emerged, at least in part, as a response to lack of replication among item writers. The purposes of this study were 1) to examine whether or not different item-writers would generate similar multiple-choice response options, and 2) to examine whether or not the response options were similar to actual student responses to the same item stems presented in open-ended format.

In answering the first purpose, two different methods of item writing were employed - a non-systematic (subject matter expertise) approach and a systematic (algorithmic) approach. Findings indicated that neither approach resulted in very similar response options being generated by writers. Although in one respect this may be viewed as evidence of item writer bias, from another point of view, this finding may be viewed as evidence of minimization of item writer bias in that the number of different response options is related to the underlying difficulty of the item.

Few options were generated by both approaches. By nature, the approaches are very different, therefore, they would not be expected to produce the same response options. However, a closer inspection of the actual options indicates that the restriction in the algorithmic approach to one error per pass contrasts with the employment of judgment by subject matter experts.

With regard to the second purpose of the study, student responses to open-ended items were compared to the responses generated by the two samples of item writers. Few response options produced by either approach corresponded with the incorrect student responses. Item-writer response options which did correspond to the student responses, were rarely the most frequently generated distracters. One reason for the lack of congruence between item-writer generated options and student-responses was the variety of incorrect student responses.

Such variation in incorrect student responses makes it apparent that the use of multiple-choice items to test basic mathematics operations will not account for all student errors. Therefore, it is likely that scores on multiple-choice tests would be inflated over scores on tests utilizing the same items in open-ended format. Students who might work problems incorrectly in completion format, would in a multiple-choice test be signalled to rework the problem (or guess) by the absence of an option matching their incorrect work.

This final assumption is yet to be tested empirically. However, the present study provides initial data leading to such a conclusion, and highlights some of the problems likely to be encountered in attempting to construct "unbiased" multiple-choice tests.

References

Berk, R. A. The application of structural facet theory to achievement test construction. Educational Research Quarterly, 1978, 3, 63-72.

Bormuth, J. R. On the theory of achievement test items. Chicago, Ill.: University of Chicago Press, 1970.

Durnin, J. and Scandura, J. M. An algorthmic approach to assessing behavioral potential: Comparison with item forms and hierarchical technologies. Journal of Educational Psychology, 1973, 65, 262-272.

Hively, W., Maxwell, G., Sension, D., and Lundin, S. Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST project. Los Angeles: Center for the Study of Evaluation, UCLA, 1973.

Popham, W. J. Educational evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Roid, G. and Haladyna, T. A comparison of several linguistic-based, multiple-choice item writing algorithms. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Roid, G. and Haladyna, T. The emergence of an item-writing technology. Review of Educational Research, 1980, 50, 293-314.

Tiemann, P. W. and Markle, S. M. Analyzing instructional content: A guide to instruction and evaluation. Champaign, Ill.: Stipes Publishing Co., 1978.

Table 1

Number and Percentage of Different Response Options Generated by Non-Systematic and Systematic Approaches

	Approach to the Generation of Response Options			
	Subject Matter Expertise		Algorithmic Approach	
ITEM	(N different/N written)	percent different	(N different/N written)	percent different
1	(7/12)	58%	(10/12)	83%
2	(8/12)	66%	(10/12)	83%
3	(7/12)	58%	(11/12)	92%
4	(7/12)	58%	(8/12)	66%
5	(8/12)	66%	(9/12)	75%
6	(6/11)	55%	(6/12)	50%
7	(8/12)	66%	(8/12)	66%
8	(6/12)	50%	(8/12)	66%
9	(8/12)	66%	(7/12)	58%
10	(8/12)	66%	(9/12)	75%
11	(9/12)	75%	(10/12)	83%
12	(10/11)	91%	(11/12)	92%
13	(7/12)	58%	(9/12)	75%
14	(10/12)	83%	(11/12)	92%
15	(10/12)	83%	(10/12)	83%
16	(9/10)	90%	(9/10)	75%
17	(10/12)	83%	(11/12)	92%
18	(9/10)	90%	(11/12)	92%
19	(8/12)	66%	(11/12)	92%
20	(11/12)	92%	(12/12)	100%
TOTAL	(166/234)	71%	(191/240)	80%

Table 2

Number Each of Different Response Options Generated by Non-Systematic and Systematic Approaches*

	Approach to Generation of Response Options																							
	Subject Matter Expertise (Number of Different Writers Deriving Option)												Algorithmic Approach (Number of Different Response Options Derived)											
Item	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
1	4	2	2	1	1	1	1						2	2	1	1	1	1	1	1	1	1		
2	3	3	1	1	1	1	1	1					2	2	1	1	1	1	1	1	1	1		
3	3	3	2	1	1	1	1						2	1	1	1	1	1	1	1	1	1	1	
4	3	3	2	1	1	1	1						4	2	1	1	1	1	1	1				
5	3	2	2	1	1	1	1	1					2	2	2	1	1	1	1	1	1			
6	4	3	1	1	1	1							4	3	2	1	1	1						
7	3	3	1	1	1	1	1	1					2	2	2	2	1	1	1	1				
8	3	3	3	1	1	1							3	2	2	1	1	1	1	1				
9	4	2	1	1	1	1	1	1					3	2	2	1	1	1	1					
10	3	3	1	1	1	1	1	1					2	2	2	1	1	1	1	1	1			
11	3	2	1	1	1	1	1	1	1				2	2	1	1	1	1	1	1	1	1		
12	2	2	1	1	1	1	1	1	1	1			2	1	1	1	1	1	1	1	1	1	1	
13	3	2	2	2	1	1	1						3	2	1	1	1	1	1	1	1			
14	2	2	1	1	1	1	1	1	1	1			2	1	1	1	1	1	1	1	1	1	1	
15	2	2	1	1	1	1	1	1	1	1			2	2	1	1	1	1	1	1	1	1		
16	2	1	1	1	1	1	1	1	1				2	2	2	1	1	1	1	1	1			
17	2	2	1	1	1	1	1	1	1	1			2	1	1	1	1	1	1	1	1	1	1	
18	2	1	1	1	1	1	1	1	1				2	1	1	1	1	1	1	1	1	1	1	
19	3	3	1	1	1	1	1	1					2	1	1	1	1	1	1	1	1	1	1	
20	2	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1

*Numbered response options do not necessarily correspond across approaches.

Table 3

Response Options Generated by Writers of Both Approaches and Number of Writers for Each Approach Generating Option

Item	Options Generated by Writers of Both Approaches	Number of Subject Matter Experts Generating Option	Number of Algorithmic Approach Writers Generating Option
1	1	1	2
	2	4	1
2	1	3	2
	2	1	1
3	1	3	1
	2	1	1
	3	1	1
4	1	3	1
	2	2	2
	3	1	1
5	1	3	1
	2	2	2
	3	1	2
6	1	1	1
	2	4	3
	3	3	4
7	1	3	1
	2	3	2
	3	1	2
	4	1	1
8	1	1	1
	2	1	3
	3	3	2
	4	3	2
	5	3	1
9	1	1	1
	2	1	3
	3	1	2
	4	4	2
	5	1	1
10	1	1	2
	2	3	1
	3	3	1
	4	1	2
11	1	1	1
	2	3	2
	3	2	1
12	1	1	1
	2	1	1
	3	1	1
13	1	3	3
14	1	2	1
	2	1	2
15	1	1	1
	2	1	1
	3	2	2
16	1	1	1
	2	1	2
	3	1	1
	4	2	2

Table 3 (continued)

Response Options Generated by Writers of Both Approaches and Number of Writers for Each Approach Generating Option

Item	Options Generated by Writers of Both Approaches	Number of Subject Matter Experts Generating Option	Number of Algorithmic Approach Writers Generating Option
17	0	-	-
18	1	1	1
	2	1	1
19	1	3	2
20	0	-	-

Table 4

Number and Percentage of Incorrect Student Responses
Matching Distracters Generated by Non-Systematic and Systematic Approaches

ITEM	NUMBER INCORRECT RESPONSES	Approach to the Generation of Response Options			
		Subject Matter Expertise		Algorithmic Approach	
		N	%	N	%
1	12	4	33%	7	58%
2	25	1	4%	6	24%
3	18	2	11%	9	5%
4	17	5	29%	7	41%
5	52	0	--	4	8%
6	13	1	8%	2	15%
7	26	2	8%	5	19%
8	38	19	50%	17	45%
9	31	12	39%	12	39%
10	31	14	45%	13	42%
11	20	3	15%	1	5%
12	27	5	19%	11	41%
13	69	3	4%	8	12%
14	61	9	15%	9	15%
15	71	5	7%	7	10%
16	34	8	24%	10	29%
17	56	1	2%	18	32%
18	56	2	4%	4	7%
19	72	12	17%	7	10%
20	92	1	1%	5	5%
TOTAL	821	109	13%	162	20%

Table 5

Number and Percentage of Different Responses
Given by Examinees Responding Incorrectly to Completion Items
N=131

Item	N Different	N Incorrect	%	N Answering	%
1	8	12	67	12	67
2	21	25	84	25	84
3	14	18	78	18	78
4	13	17	76	16	81
5	38	52	73	49	78
6	11	13	85	11	100
7	19	26	73	25	76
8	23	38	61	37	62
9	20	31	65	28	71
10	17	31	55	29	59
11	18	20	90	19	95
12	18	27	67	25	72
13	45	69	65	65	69
14	50	61	82	55	91
15	56	71	79	63	89
16	18	34	53	25	72
17	29	56	52	44	66
18	22	56	39	24	92
19	43	72	60	49	88
20	44	92	48	53	83
TOTAL	527	821	64	672	78