DOCUMENT RESUME

ED 036 870 24 CG 005 169

AUTHOR Baker, Eva L.
TITLE The Effects of Manipulated Item Writing Constraints on the Homogeneity of Test Items. Center for the Study of Evaluation Reprint Series No. 11.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO BR-6-1646
PUB DATE Mar 70
CONTRACT OEC-4-6-01646-1909
NOTE 12p.; Paper presented at National Council for Measurement in Education Conference, Minneapolis, Minnesota, March, 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$0.70
DESCRIPTORS *Behavioral Objectives, Educational Objectives, *Item Analysis, Objectives, *Research Methodology, *Test Construction, Testing, *Tests, Test Selection

ABSTRACT

To help teachers who must produce test items to measure instructional objectives, 54 teacher education candidates participated in an experiment where easily understood constraints on item production were manipulated. Four forms of a test item writing exercise sheet were randomly distributed, each asking for the production of eight sample test items, two for each specified topic. The subjects produced 16 items, to be used for seventh grade students. Two 16 item tests were constituted, one on subtraction and one on current events. The tests were administered to 51 junior high school students. Means and standard deviations of the items were computed, and analysis of variance for the subtest means was conducted for each replication. Significant differences ($F=8.3$, $df=3, 12$) were observed for subtraction. For the current events data differences were not significant. Findings are limited by the number of items on each subtest. Further staff studied are investigating how tc produce items truly congruent with objectives and how best to translate these findings into practical procedures for teacher. (Author/CJ)

THE EFFECTS OF MANIPULATED ITEM WRITING CONSTRAINTS ON THE HOMOGENEITY OF TEST ITEMS

Reprint #11

UCLA
Graduate School of Education
Los Angeles, California

Marvin C. Alkin
Director

UCLA Graduate School of Education

The CENTER FOR THE STUDY OF EVALUATION (CSE) is one of nine centers for educational research and development, sponsored by the United States Department of Health, Education, and Welfare, Office of Education. Established at UCLA in June, 1966, CSE is devoted exclusively to finding new theories and methods of analyzing educational systems and programs and gauging their effects.

The Center serves its unique function with an interdisciplinary staff whose specialties combine for a broad, versatile approach to the complex problems of evaluation. Study projects are conducted in three major program areas: Evaluation of Instructional Programs, Evaluation of Educational Systems, and Evaluation Methodology and Services.

Paper to be Presented at the Annual Conference of the
National Council for Measurement in Education
Minneapolis, Minnesota, March, 1970

THE EFFECTS OF MANIPULATED ITEM WRITING CONSTRAINTS ON THE HOMOGENEITY OF TEST ITEMS[1]

Eva L. Baker
Graduate School of Education
University of California, Los Angeles

The greatest curse, wise men have said, may be to have your wishes come true. A case in point is the advocacy of objectives-based instruction and evaluation, where teachers test, teach, and retest children until desired levels of mastery are reached. The tests used in this type of instruction differ from commercially produced achievement tests because they are directed toward specific program goals, usually stated in operational language. Program planning and budgeting systems are expanding the appeal of such approaches, and the call for objectives and items has increased. While a fledgling institution[2] has emerged to bear part of the burden for generating some of the objectives and items needed for large scale implementation of such an approach, it has become clear that more items will be demanded than can currently be prepared. Obviously, if a teacher needs a great number of items for iterative testing, he will either produce them himself, or go without and revert to a more usual instructional pattern.

What kind of help can be provided for the teacher who must produce test items to measure his instructional objectives? Do simple procedures exist which allow the teacher to produce homogeneous test items? Some clear alternatives to control item production involve the use of behaviorally stated objectives, sample test items and simplified item forms.

Improved production of test items has historically been one of the benefits emphasized by curriculum specialists advocating behavioral objectives. Broadly stated objectives make the estimate of congruence between objective and item difficult to determine. For example, if one were asked to produce items to measure an objective such as "understanding of statistical concepts", a great number of items would be considered suitable, and depending upon which set happened to be used by the instructor, vastly different notions about student achievement would be inferred. However, if the objective was modified to "the student would have to select and justify a statistical analysis for

---

[1] The research herein reported was partially supported by the Center for the Study of Evaluation, UCLA, pursuant to a contract with the U.S. Office of Education, Department of Health, Education and Welfare, under the provisions of the Cooperative Research Program.

[2] The Instructional Objectives Exchange, founded with the support of the UCLA Center for the Study of Evaluation.

those research designs described by Campbell and Stanley," performance on an appropriate set of items should give a fairly good idea of the attainment of the objective. A further way to reduce the heterogeneity of responses to the items might be to employ a standard format for each item. Additionally, if the content to be sampled was made more precise, then one would assume that increased homogeneity would be demonstrated by sets of items measuring the same objective.

The item form, under development at Minnesota,[3] describes both the format which the items in a set should take and the content limits which should be observed. Attention has been directed to variants of this idea both at UCLA and the Southwest Regional Laboratory for Educational Research and Development.[4] The Project for Research on Objectives-Based Evaluation (PROBE), a program of the UCLA Center for the Study of Evaluation, used generation rules for producing sets of items to accompany objectives for the Instructional Objectives Exchange. These rules limited the format of the item and defined the content area to be assessed. However, when teachers were asked to use these generation rules to produce additional items, to measure the objectives, they were appalled by the difficulty they experienced in deciphering the technical language of the rules.

Method

To gain a modest amount of additional information, an experiment was conducted where various easy-to-understand constraints on item production were manipulated for a population of teacher educational candidates. Effects on item homogeneity were to be observed.

Subjects. Fifty teacher education candidates enrolled in a curriculum course were the subjects who generated the test items. These students were seniors and graduates enrolled in summer session. They were given an ostensible test writing exercise as one assignment in their course.

Treatments. Four forms of a test item writing exercise sheet were randomly distributed to the subjects. Each form asked for the production of eight sample test items, two for each of the following topics: Current Events, Subtraction, Graphs, and Punctuation Errors. Form one of the exercise provided an objective stated very generally. For the first topic, the statement was as follows: "Awareness of the relationship of personalities to current events." Form two provided a behavioral objective to guide the item writing. The objective for the Current Events topic was: "To be able to identify people associated with important current events." Even though considerable clarity is reflected in this objective, a number of interpretations of it were

---

[3] See papers presented at NCME symposium on Criterion-Referenced Measurement, 1970.

[4] Baker, R. L., Gerlach, V. S., Schutz, R. E. and Sullivan, H. J., "Developing Instructional Specifications", in Developing Instructional Products, W. James Popham (Editor), Southwest Regional Laboratory for Educational Research and Development, Inglewood, California, 1968.

obviously possible. For the same topic, Form three again listed the objective, but, in addition, supplied a sample multiple choice item in which the current event was stated in the stem and alternatives were the names of personalities. This condition is identical to the way in which objectives from the Instructional Objectives Exchange are disseminated, since each objective is accompanied by a sample item. Form four also included the same behavioral objective for this topic. In addition, five statements designed to constrain the type of item produced were provided as follows:

a. The format should be multiple choice.

b. There should be only four alternatives provided.

c. The current event description should appear in the stem of the item; people's names should form the alternatives.

d. Only one answer should be right for each question; "none of the above" or "all of the above" should not be alternatives.

e. Current events should be limited to occurrences within the last two years which probably received front page space in the newspaper. An example might be space exploration.

The first four statements related only to the format of the item while the last statement attempted to restrict the content domain from which the item writer could draw. The sample multiple choice item provided in Form three was an instance of an item which would fit the description given in Form four.

Procedure. The subjects were allowed approximately 90 minutes to produce the 16 items. Directions were given in each exercise form that the items would be used for seventh grade students. Subjects were asked to avoid inflated language, provide necessary test directions, and to supply either the right answer or criteria for judging each answer, so items could be scored.

Composition of the Test. Items produced were segregated by treatment and by topic. Two 16 item tests were constituted, each composed of four items randomly selected from those produced by item writers in each treatment, one for the topic of subtraction and one for current events. Within each topic the items were randomly ordered except all constructed responses were grouped together to minimize the distraction of changing response sets. The topic of subtraction was selected because performance in that area might simulate that of an "instructed" group, since subtraction practice has generally been encountered by most seventh grade students. Current events, however, might represent an area given less systematic instructional attention. Perhaps differing levels of competence for the topics might be reflected in the data.

Field Trial. Fifty-one seventh grade students in a Los Angeles junior high school were administered the 32 items. Children were told that they were being compared with other seventh grade students in their subtraction and current events skills and were given one hour to complete all 32 items. Right

answers were read to them by their teacher after the entire test had been completed.

Data Analysis and Results

Means and standard deviations of the items were computed and are reported in Table 1. Analysis of variance for the subtest means was conducted for each replication. Significant differences ($F=3.3$, $df=3, 12$) were observed for the subtraction topic. Items produced under the most constrained conditions, that is, with a sample test item as a model or the modified item form, produced items with higher means. The same order effect was observed in the current events data but the differences were not found to be significant.

On a common sense basis, one would generally assume that items generated under a given treatment condition would correlate better among themselves than with subtests produced under different treatment conditions. However, an exception might be found for those items produced under the nonbehavioral objective condition. Such items might be expected to differ considerably from one another and might fail to correlate highly with each other or with any of the other subtests.

Point biserial correlations were computed for each subtest generated by the four treatments for both replications (See Tables 2 and 3). The average correlation of items with their own subtest was compared with the average correlation of items with each of the other three subtests. Four separate analyses of variance were conducted for the two topics. For the current events topic, significant differences found for each of the "constrained" treatments, that is, items produced with either an objective, test item, or modified item form as a guide, tended to correlate better among themselves than with items produced by the other treatments. The exception, in current events, was the analysis conducted on the nonbehavioral subtest. No significant differences were obtained, and in fact, none of the mean correlations was above .35. In the subtraction replication, significant differences were found on each of the analyses of variance conducted. Perhaps because the topic of subtraction, in itself, provided sufficient structure, the correlations observed were considerably higher.

Implications

Modest evidence was found that items produced with some constraints were more homogeneous than items produced under general conditions for the current events topic. The different treatments did seem to have predicted effects in both the replications. The disconfirming evidence, the significant differences found in the subtraction replication for the nonbehavioral treatment, might be a function of the precision of the subject matter itself. When one inspects the mean correlations of the "constrained" treatments, no particular advantage was found for either the behavioral objective, sample test item, or modified item form. In the current events replications, the correlations produced by these treatments are within one point of each other.

on the subtraction replication, they are within four points of each other.

One factor which obviously limits the findings of this study was the number of items on each subtest. The selection of four items for each subtest was not divinely inspired. Rather, the number of items selected was in part determined by the original effects of the treatments. Subjects in the treatment one, writing test items under the "nonbehavioral" condition tended to generalize the lack of structure to the extent that only four items of the 26 produced for the topic of current events were scorable, that is, included either right answers or means for determining the right answer. One of these items was in multiple choice format while the other three were completion items. So the usable items generated by the treatment contained much more structure than most of the items produced by subjects in that treatment group. One could expect even more variability than was observed to be associated with the disparate items which were generated but not usable, e.g., "Write and essay describing the contribution of a famous 20th century man." Even fewer usable items were produced on the topics of punctuation errors and graphs.

Clearly, the study did not produce evidence compelling enough to change the current method of providing teachers with a sample test item accompanying each objective. Further studies are underway by the PROBE staff continuing to investigate how to produce items truly congruent with objectives and how one can best translate these findings into practical procedures for teachers.

Table 1. Means and Standard Deviations of Items by Treatment Group

	Item	Subtest 1 (Nonbehavioral) $\bar{X}$	s	Item	Subtest 2 (Objective Only) $\bar{X}$	s	Item	Subtest 3 (Objective Plus Item) $\bar{X}$	s	Item	Subtest 4 (Simplified Item Form) $\bar{X}$	s
CURRENT EVENTS	1	.02	.14	5	.33	.47	9	.49	.50	13	.31	.46
	2	.65	.48	6	.67	.47	10	.73	.45	14	.39	.49
	3	.00	.00	7	.33	.47	11	.65	.48	15	.90	.30
	4	.73	.45	8	.53	.49	12	.41	.49	16	.37	.48
	Subtest	.35	.39		.47	.16		.57	.14		.50	.27
SUBTRACTION	1	.61	.49	5	.82	.38	9	.86	.19	13	.98	.14
	2	.65	.48	6	.65	.48	10	.86	.34	14	.94	.23
	3	.63	.48	7	.80	.40	11	.96	.19	15	.92	.27
	4	.38	.32	8	.80	.40	12	.96	.19	16	.86	.34
	Subtest	.69	.12		.77	.08		.94	.05		.93	.05

Table 2. Mean Point Biserial Correlations for Subtest Items with Total Subtests

CURRENT EVENTS

	Total Subtest 1	Total Subtest 2	Total Subtest 3	Total Subtest 4
Items in Subtest 1	.34	.06	.09	.05
Items in Subtest 2	.06	.56	.20	.21
Items in Subtest 3	.08	.20	.57	.22
Items in Subtest 4	-.003	.21	.20	.56

Table 3. Mean Point Biserial Correlations for Subtest Items with Total Subtests

	SUBTRACTION			
	Total Subtest 1	Total Subtest 2	Total Subtest 3	Total Subtest 4
Items in Subtest 1	.60	.24	.14	.07
Items in Subtest 2	.25	.62	.13	.13
Items in Subtest 3	.16	.17	.65	.43
Items in Subtest 4	.11	.11	.40	.61

This publication is issued pursuant to a contract with the Bureau of Research, Office of Education, U.S. Department of Health, Education and Welfare. The views and findings expressed herein are not necessarily those of the U.S. Office of Education, and no endorsement is intended or implied.