DOCUMENT RESUME

ED 141 395 TM 006 344

AUTHOR Tenenbaum, Arlene Bonnie; Miller, Christine A.
TITLE The Use of Congruence between the Items in a Norm-referenced Test and the Content in Compensatory Education Curricula in the Evaluation of Achievement Gains.
PUB DATE [Apr 77]
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977) ; Table 2 may be of marginal legibility due to small type

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Gains; Achievement Tests; *Compensatory Education Programs; *Content Analysis; *Curriculum Evaluation; Elementary Secondary Education; Evaluation Criteria; Evaluation Methods; Field Studies; *Item Analysis; Learning Modules; *Norm Referenced Tests; Standardized Tests; *Test Validity
IDENTIFIERS Metropolitan Achievement Tests

ABSTRACT

In the evaluation of Project Information Packages (PIPs), a content analysis was performed to detect congruence between items in a norm-referenced test and the content in six exemplary compensatory education program curricula. Gains on congruent items were used to assess the effectiveness of the programs. Preliminary results show that the amount of congruence was too small to make strong inferences, but that gains on congruent items were slightly higher in well-implemented programs. The procedures can be easily replicated for evaluations which require that gains on norm-referenced tests be the major criteria for success. (Author)

The Use of Congruence between the Items in a Norm-referenced Test and the Content in Compensatory Education Curricula in the Evaluation of Achievement Gains *

ARLENE BONNIE TENENBAUM **
CHRISTINE A. MILLER

Stanford Research Institute

Presented at the American Educational Research Association Meetings
April 1977 (22.05)

*Please do not cite the results in this paper until contractual obligations have been fulfilled (expected March, 1977).

**Arlene Tenenbaum was a consultant to SRI during the conduct of this study. Inquiries may be sent to the first author at 213 West Portola Avenue, Los Altos, California, 94022.

ABSTRACT

In the evaluation of Project Information Packages (PIPs) the content validity of the Metropolitan Achievement Test was examined by searching for congruence among test items, the content in the project curricula, and the curricula taught to the pupils. Patterns of achievement were used as descriptive indices of program effectiveness. A model analysis design is presented for replication in other evaluations which require gains on norm-referenced tests to be the major criteria for program success.

BACKGROUND

A methodology was developed to establish congruence between the knowledge and skills tested by a norm-referenced achievement test (MAT '70) and the content in the curricula specified in six, exemplary, compensatory education programs. While it was pursued within the evaluation of the field test of Project Information Packages (PIPs), the methodology bears directly upon a familiar dilemma in evaluation of field studies.

In many specially funded programs an evaluator is obliged to show gains on a norm-referenced achievement test as evidence of success. Gains can only be expected if the content in the test matches the curriculum taught in the program. This expectation follows from a fundamental assumption in constructing an achievement test, namely content validity. The content, in toto, while representative of a spectrum of achievement in classrooms across the U.S., may not be valid for pupils in a special program. Instead, only a subset of the test items might actually have been taught in a particular program. Logically, then, gains on the total score do not represent a reasonable measure of the particular program.

In the first phase of this evaluation few gains were found at any grade for the six programs (Stearns). One possible explanation for this failure could be the lack of content validity of the norm-referenced test.* A search for congruence would test this possibility.

A methodology to establish congruence would require two procedures: one to describe the content in the test and the other to determine what content was taught. Few systematic studies of the congruence between test items and curricula have been conducted. In an earlier study congruence between a norm-referenced test (Coop. Prim.) and state texts (Ca.) for first grade vocabulary was reported at only 55% (by permission, Bianchini). For vocabulary the researcher could search for exact matches; that strategy would

*Since content validity may also be related to differences between the norm group and the group that was tested, the avenue was explored in the evaluation of PIPs, too.

not work for all of the content in this test battery, especially reading comprehension and math problem-solving. In this study we also looked at the content outlines provided by the publisher; they were not sufficiently precise. It was necessary to formulate our own criteria and descriptions of the content in each item.

When the new science curricula was introduced, teachers reviewed the college board achievement tests and marked the items covered in their courses. Some part-whole scales were then provided to show congruence for the new and traditional curricula (eg. Malcolm and Watkins). That method was rejected for our study, because we sought "naturalistic" data, ie., a source of data typically found in a particular setting and unbiased by foreknowledge of the purpose of an eventual collection. The project teachers were requested to turn in their daily lesson plans and we did not explain the rationale for the collection. These records were used to determine what was taught.

METHOD

Congruence: The following tasks were completed to establish congruence: (1) rules were formulated to describe the knowledge and skills in each item in each subtest; (2) the curricula in the projects was searched for the presence of these rules and appropriate content keyed to each item; (3) the curricula entered on a daily record for each pupil was listed; and (4) when the curricula taught to each pupil overlapped with the curricula keyed to a test item, a match was declared.

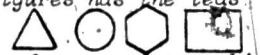
Rules: The rules for each item were formulated to satisfy two criteria. First, each item was viewed as a single, independent skill and as many discrete features of an item as we could determine were specified. Second, a strategy of "near transfer" was adopted. All of the features had to be represented in the curricula exactly as they were found in the test format. It seemed likely that the content in the word knowledge and arithmetic computation items could be found reproduced almost exactly in the curriculum, as they were in the earlier study cited (Bianchini); the only variation might occur in the actual values of the numbers in math computation.

The test publishers posed the item questions in the reading

comprehension test quite consistently, and rules were constructed for matching those items with the content. Since the questions are associated with a reading passage, rules for the passage as well as rules for the kind of question the item posed were linked and ranges for matches were established. Existing quantitative rating systems, such as the Dale-Chall readability formula and average sentence length, were selected. Such measures are unquestionably rough indices of comprehension levels, but they can be computed clearly.* Thus, the rule for items in the reading comprehension test specified the type of item (main idea, literal, inferential) and features of the associated passage.

For a small percentage of the items no verifiable rule could be constructed. The following rule illustrates the application of these criteria and the type of description for one category of math story problem.

ITEM: *Which of these figures has the least number of line segments?*



RULE: *In order to respond correctly to this item a pupil must have been taught*

(a) recognition of planar shapes: triangle, rectangle, circle, and hexagon

(a1) same scale

(b) concepts: counting, least number, line segments

(c) vocabulary (sight recognition and understanding): figures, least, number, line, segments

Pupil records: From the daily lesson plans of the teachers usable data was extracted; incomplete, illegible entries were tabulated and used only to keep track of the sample size of usable data relative to the entire collection. The use of the remaining content was then verified; in reading programs, for instance, in which aides read aloud to pupils, it was ascertained that pupils also read silently to themselves, as would be required in the test setting. Since documents, like lesson plans, might not reveal the degree to which a skill was taught, criteria were set to compensate for variations among teachers as well

*In various studies the difficulty index derived from the Dale-Chall formula has accounted for as much as 50% of the criterial variance.

as standardizing the impact of curricula across test items. A particular word was considered taught, for example, if the records showed a pupil was exposed twice to curricula that contained the word in a well-marked exercise. Other heuristics were developed for content in reading and arithmetic.

Much of the effort in the analysis of the curriculum was absorbed by fragmenting the curricula into units that would match single items on the test. Sometimes one booklet covered only one skill, while in another series different skills appeared on the same page. In the former, an entry was listed to mark the beginning and ending pages, whereas, in the latter, each problem was noted. Thus, a file was built, showing the recommended curricula taught to each pupil. It was merged with another file, showing which curricula was keyed to the test items.

Analysis: For those pupils who took the same test in the fall and spring patterns of achievement (fail-pass, pass-pass, pass-fail, fail-fail) were tallied to compare performance on congruent and non-congruent items. A model factorial design was later devised to incorporate the variables which appear to influence the patterns of achievement.

RESULTS

As of this writing, only preliminary results are available.* The amount of congruence (item x content x pupil) appears to be very small, 5-20%, and decreases with grade level, except in the subtest arithmetic computation. Some gains appear for that test.

The results suggest that the following factors could be included in a model design: (1) Between-subjects: project, school quality, type of materials, teacher-pupil arrangement, and grade; (2) Within-subjects: subtest, congruence of item. The univariate dependent measure would be the pattern of achievement (Another variate, item difficulty, could be added for multivariate analysis. Alternatively, achievement patterns could be split so that entering performance (fall) could be viewed as an independent variable.) Possible levels for these factors will be

*Sample patterns will be available for display at the session.

discussed at the session.

EDUCATIONAL IMPLICATIONS

This method for establishing congruence confirms our expectation; on some subtests pupils learned and, therefore, gained slightly on some items which they were taught. These procedures can be easily replicated for evaluation of similar programs, either national or local, in which norm-referenced tests are required and used as the main criteria for success.

Certainly, some educators and lay persons will not accept evidence of success which rests upon a small percentage of test items, regardless of the amount of gains within that percentage. When technical constraints are satisfied and issues of use and interpretation resolved, size itself becomes a philosophical preference. We can only speculate about the causes for the lopsided congruence/non-congruence in this project. Perhaps, instructional guidance in the packages was weak or the level of pupil performance and complementary curricula lay far below the range on the test. Hopefully, replications of this methodology would reveal differences clearly attributable to some factors in the model design.

Further analyses to compare the test-retest reliability of performance between congruence and non-congruent items would lend stronger support to the methodology when the patterns of achievement are different and congruence/non-congruence lopsided. At the same time, this next step could address the question of success with reference to growth nationally, the purpose, after all, of using a norm-referenced test in an evaluation.

APPENDIX

Table 1

THE NUMBER OF MAT ITEMS CONGRUENT WITH PIP-SPECIFIED MATERIALS AT THE FOURTH AND EIGHTH GRADES

TOTAL READING		TOTAL MATHEMATICS		
Word Knowledge	Reading Comprehension	Computation	Concepts	Problem-Solving
Fourth Grade				
T = 50	T = 45	T = 40	T = 40	T = 35
N = 28	N = 17	N = 40	N = 7	N = 6
N = 56% of T	N = 37.7% of T	N = 100% of T	N = 17.5 % of T	N = 17.1% of T
Eighth Grade				
T = 50	T = 45	T = 40	T = 40	T = 35
N = 2	N = 3	N = 28	N = 3	N = 5
N = 40% of T	N = 6.6% of T	N = 70% of T	N = 7.5% of T	N = 14.2% of T

Key:

T = Total number of items in MAT subtest

N = Number of items known to be congruent

Table 2

COMPARISON OF ACHIEVEMENT PATTERNS FOR SAMPLED PUPILS

Name of Test	Grade	GOOD TEACHER								POOR TEACHER							
		Congruent				Noncongruent				Congruent				Noncongruent			
		FP	PP	FF	PF	FP	PP	FF	PF	FP	PP	FF	PF	FP	PP	FF	PF
Word Knowledge	4	2	8	1	0	109	73	189	68								
		.18	.73	.09	0	.25	.17	.43	.15								
		2	[1]	3	4	2	3	(1)	4								
	8	(no congruent response patterns)															
Reading Comprehension	4	45	133	81	27	321	286	614	242	19	23	15	3	109	76	134	76
		.16	.47	.28	.09	.22	.20	.42	.16	.32	.38	.25	.05	.28	.19	.34	.19
		3	[1]	2	4	2	3	(1)	4	2	[1]	3	4	2	3-4	(1)	3-4
	8	(only 3 congruent response patterns)															
Math Computation	4	(all items congruent-no comparison of response patterns possible)															
	8									30	50	27	11	181	388	371	143
										.25	.43	.23	.09	.17	.36	.34	.13
										2	[1]	3	4	3	(1-2)	(1-2)	4
Math Concepts	4	(no congruent response patterns)															
	8									2	8	2	1	62	156	152	56
										.15	.62	.15	.08	.14	.37	.36	.13
										2-3	[1]	2-3	4	3-4	(1-2)	(1-2)	3-4
Math Problem Solving	4	1	12	4	3	48	38	114	25	3	4	1	0	16	6	15	6
		.05	.60	.20	.15	.21	.17	.51	.11	.38	.50	.12	0	.37	.14	.35	.14
		4	[1]	2	3	2	3	(1)	4	2	[1]	3	4	(1-2)	3-4	(1-2)	3-4
	8									10	9	4	1	215	296	316	111
										.42	.37	.17	.04	.23	.31	.34	.12
										[1]	2	3	4	3	2	(1)	4

Pre-Post Response Patterns: FP (Fail/Pass); PP (Pass/Pass); FF (Fail/Fail); PF (Pass/Fail)

Line 1 = Number of specified response pattern for a subset (across all PIPs), e.g. number of FP for a good teacher/congruent items

Line 2 = $\frac{\text{Specified Response Pattern}}{\text{Total Response Patterns for Subset}}$

Line 3 = Rank order of response patterns within subset

Table 3

POSSIBLE MODELS FOR ANALYSES

I Program (child = unit of analysis)

Type: Factorial
ANOVA (all factors are categorical)

Independent Variables

Between-subjects
- Project
- School
- School quality
- Teacher-pupil arrangement
- Grade
- Materials

Within-subjects
- Subtest
- Congruence of item (for pupil)
- Materials

Dependent Variable

Patterns of Fall-Spring Pass/Fail

II Norm-Referenced (item = unit of analysis)

Type: Multiple regression (simple)
One equation for congruent and one equation for noncongruent

Independent Variables

X_1 # children with fail-pass
X_2 # children with pass-pass
X_3 # children with fail-fail
X_4 # children with pass-fail

Dependent Variable

Y_1 mean percent difficulty of item