DOCUMENT RESUME

ED 227 165 TM 830 187

AUTHOR Koffler, Stephen L.
TITLE A Longitudinal Analysis of Curricular Validity for a Minimum Competency Testing Program.
PUB DATE Apr 83
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, Canada, April 11-15, 1983). Tables 1 and 2 contain small print.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Basic Skills; Court Litigation; *Curriculum; Elementary Secondary Education; Longitudinal Studies; *Mathematics Instruction; *Minimum Competency Testing; Psychometrics; *Reading Instruction; Testing Problems; Testing Programs; *Test Validity
IDENTIFIERS Content Validity; *Curricular Validity; Debra P v Turlington; Modified Caution Index; *New Jersey Minimum Basic Skills Program

ABSTRACT

This study examined the curricular validity of the New Jersey Basic Skills test, a minimum competency test administered to all public school students in grades 3, 6, 9, and 11 to measure basic skills in reading and mathematics. Based on examinations of a Modified Caution Index, there were differences in the usual response patterns for both reading and mathematics. This result suggests that within districts there may be differences in the content coverage and emphasis placed on some of the subsets of items contained on a minimum competency test. Because differences were noted across districts and curricular programs, there is the suggestion that there may be problems with using one test. Other, more detailed non-test-based analyses should be conducted to further examine the curricular validity and also the instructional validity of this test. The present analyses provide an initial insight into possible differences between test and curricular matches. (Author/PN)

A Longitudinal Analysis of Curricular Validity For A Minimum Competency Testing Program

Stephen L. Koffler
New Jersey State Department of Education

Paper presented at the Annual Meeting of the American Educational Research Association
Montreal, Canada, April 11-15, 1983

A Longitudinal Analysis of Curricular Validity For A Minimum Competency Testing Program

Stephen L. Koffler
New Jersey State Department of Education

INTRODUCTION

Until recently a narrow definition of content validity has been used when considering achievement tests. If the items measured the test's objectives, then the test was considered to be content valid for all examinees regardless of their background, school attended, or instructional program. However, because of court decisions related to minimum competency testing and the use of such tests for high school graduation, the definition of content validity now has been broadened to include consideration of both curricular and instructional validity.[1,2]

The issues of curricular and instructional validity surfaced in the Debra P. v. Turlington case. In that case, the plaintiffs challenged Florida's 1976 high school graduation requirement law which mandated that students had to pass the Florida Functional Literacy Test (a test developed by the Department of Education), and satisfy other requirements, to receive a high school diploma.

---

[1] Curricular validity refers to the match between the skills tested and those in the curriculum; instructional validity refers to the match between the skills tested and those taught.

[2] As Madaus(1983) indicates, the definition of content validity historically has included the concepts of curricular and instructional validity. In practice, however, the narrow definition of content validity was used. For a detailed treatise on the issues of curricular and instructional validity read The Courts, Validity, and Minimum Competency Testing, edited by George F. Madaus, Boston: Kluwer-Nijhoff Publishing, 1983.

A key issue in the Debra P. litigation was whether a test used as a graduation requirement " ... should only measure that which the schooling has offered the students."(Pullin, 1983). Many of the arguments pertaining to this issue centered on the definition of content validity. The defendants argued for the narrow definition of content validity, i.e., the match between items and skills. The plaintiffs argued that content validity had to include curricular and instructional validity.

The appeals court agreed with the plaintiffs and ruled that in determining the content validity of the Florida Functional Literacy Test, the Florida Department of Education must address the issue of whether the test covered the material taught. Madaus(1983) accurately summarized the situation: "The court's decision to broaden the meaning of content validity to include evidence that pupils had been taught the materials on a certification test adds an important new dimension to the validation process. If the test is to be used as a graduation requirement, then the court is asking the state for evidence that the test is measuring things that pupils had fair opportunity to learn."

Clearly, Minimum Competency Tests(MCT) which have been developed with care, based upon rigorous professional standards, should have content validity in the narrow sense. However, the broader question is whether the MCTs, especially those used for graduation decisions, have curricular and instructional validity.

In the high schools, there are four types of curricular programs -- college preparatory/academic, general, business/commercial and vocational/industrial arts. The scope of these

programs' curricular offerings differs within and among schools. This raises an important question -- can a single state-developed Minimum Competency Test be curricular valid across all programs and all high schools? In broader terms, can state-developed tests be used fairly as a requirement for high school graduation.

This study focused on curricular validity. Its purpose was to examine the curricular validity of the New Jersey statewide minimum competency test, considering the different high school programs. The study also examined the change in the curricular validity during the five years of the program's existence.

MEASURING CURRICULAR VALIDITY

There are many methods to analyze the curricular validity of a test. Popham and Lindheim(1981) identified two methods. The first is based on an analysis of the instructional materials, including textbooks, course syllabi and teachers' lesson plans. The second involves an analysis of the interactions in the classroom. Schmidt, et.al. (1983) developed a taxonomy which enabled them to measure the content of instruction, tests and curricular materials. The taxonomy maps the test's items into its content specifications and permits one to determine the degree to which the test item taxonomy map is subsumed under the specification map. Leinhardt(1983) suggested procedures based on an analysis of the match between scope and sequence charts and test descriptions of content covered, an analysis of texts by either item or computer search, and an analysis of instruction by teacher observation.

All of these procedures as well as similar ones suggested by others are difficult to apply. They rely on the collection of

considerable data from a broad range of individuals. There are other procedures to assess curricular validity based on tests and test results from which data are more readily obtained.

Harnisch & Linn(1981) provide a comparison of techniques which can be used to identify unusual response patterns on test items. They said that an analysis of the response patterns can be used to discover relationships between specific tests and the curricula. They also suggested that differences in performance on items measuring certain skills could indicate weaknesses in the teaching of the skills in different districts.

According to Haney(1983), using tests to examine curricular validity has two disadvantages. First, the methods rely on the test data. If the validity of the test is questionable, then the use of the test results is limited. Second, the procedures are applicable only for groups of students, not individuals; however the real concern is for the individual.

Haney's(1983) limitations of the test-based procedures can be overcome. As previously indicated, content validity in the narrow sense should be assured because of the procedures and care used to develop the test. The issue of group v. individual analysis would be a more serious concern were one considering instructional validity. For the analysis of curricular validity, which is a prerequisite for an examination of instructional validity, an examination of group results will suffice. Such an analysis could provide information regarding differences in exposure to different subject matter and the manner in which that subject matter has been taught. (Harnisch & Linn, 1981). Thus, the most

practical method for an initial examination of curricular validity is based on the test results.

BACKGROUND/DATA SOURCE

The New Jersey Minimum Basic Skills Tests(MBS) have been administered annually since spring 1978 to all public school students in grades 3, 6, 9 and 11. These tests measure reading and mathematics minimum basic skills which people in New Jersey determined were the skills students must master, at a minimum, by spring of the tested grades. In 1979 a state law was passed which established uniform statewide high school graduation requirements. Beginning with the ninth grade class in 1981-1982, students have to meet certain curricular and attendance requirements and also have to pass the ninth grade statewide test to obtain a high school diploma.

Each test contains approximately 100 four-option multiple choice items and takes 90 minutes to complete. All items are rewritten each year although the skills upon which the tests are based remain the same. An equating procedure assures the equivalence of scores across each year's forms and a uniform score scale (0-100) makes consistent the reporting of the results. Finally, in addition to reporting total test scores, scores are reported for three reading subskill 'clusters' (word recognition, reading comprehension and study skills) and four mathematics subskill 'clusters' (computation, number concepts, measurement & geometry, and problem solving & applications).

For the present research five school districts, representing each of the five major types of school districts in New Jersey

(urban, suburban, rural, regional and vocational), were randomly selected. Ninth grade students' results in those districts were used because of that grade's relationship to the graduation law. Data were obtained for the ninth grade students in the first (1978), third (1980) and fifth (1982) year of the MBS program.

The final data element collected was the students' high school program. Each year the ninth grade students were asked a series of background/contextual questions. One such question asked: "Which of the following best describes your present high school program?" The possible responses were limited to -- business/commercial, college preparatory/academic, vocational/industrial arts, and general. This information and the students' test results were used to examine the curricular validity of the MBS.

METHODOLOGY

Harnisch & Linn (1981) compared eight different indices designed to determine whether an individual's pattern of responses on an achievement test was unusual.

> Items which are generally difficult for most students may be relatively easy for students who have been in classes where that particular content was emphasized. Such variation from the norm may lead to the systematic over- or under- estimation of an individual's or group's level of achievement, distorting the measurement results.
>
> These indices could be used to identify individuals for whom the standard interpretation of the test score is misleading, or identify groups with atypical instructional and/or experiential histories that alter the relative difficulty ordering of the items. In addition, the items that contribute most to high values on an index for particular subgroups could be identified and judgments made regarding the appropriateness of the item content for those subgroups. (Harnisch & Linn, 1981).

A Modified Caution Index (C_i^*) will be used for this study. Harnisch & Linn(1981) concluded that C_i^* was the best index to use to examine unusual response patterns because it was the least correlated to total test score of the eight indices they compared.

DESCRIPTION OF CAUTION INDICES

Sato developed a matrix called the Student - Problem(S-P) Table to define an index of the degree to which an individual's response pattern is unusual. (See Tatsuoka 1978). Each row of the matrix represents an examinee while each column represents an item. Cell entries are either ones for correct responses or zeros for incorrect responses. The columns of the matrix are arranged from left to right in ascending order of item difficulty; the rows are arranged from top to bottom in descending order of total number of correct answers.

If the items on a test formed a perfect Guttman Scale (Guttman, 1941) the S-P Table would consist of all ones in the upper left corner and all zeros in the lower right corner. Anyone who responded correctly to a difficult item would have answered all easier items correctly. There would be no unusual response patterns because everyone with a given total score would have the same response pattern. However, because perfect Guttman Scales are unlikely on achievement tests, a typical S-P Table will be characterized by mostly (but not all) ones in the upper left corner and mostly (but not all) zeros in the lower right corner.

Sato(1975) developed an index based on the S-P Table called the Caution Index (C_i). C_i provides information about an examinee which is not contained in the total score. Examinees with large

values for C_i have unusual response patterns. Harnisch & Linn(1981) suggest that "unusual response patterns may result from guessing, carelessness, high anxiety, an unusual instructional history or other experiential background, a localized misunderstanding that influences responses to a subset of items, or copying a neighbor's answers to certain questions." Thus, those students' test score should be interpreted with caution

Sato's Caution Index for the i^{th} examinee is as follows:

$$C_i = \frac{\sum_{j=1}^{n_{i.}} (1 - u_{iJ}) n_{.J} - \sum_{j=n_{i.}+1}^{J} u_{iJ} n_{.J}}{\sum_{j=1}^{n_{i.}} n_{.J} - n_{i.} \left(\sum_{j=1}^{n_{i.}} n_{.J} / J \right)} \qquad (1)$$

where

$i = 1, 2, \ldots, I$ indexes each of the I examinees;

$J = 1, 2, \ldots, J$ indexes each of the J items;

$u_{ij} = \begin{cases} 1 \text{ if examinee } i \text{ answers item } J \text{ correctly,} \\ 0 \text{ of examinee } i \text{ answers item } J \text{ incorrectly,} \end{cases}$

$n_{i.}$ = number correct for the i^{th} examinee,

$n_{.J}$ = number of correct responses to the J^{th} item.

The problem with C_i is that large values may occur, especially in cases where a very high scoring examinee misses one easy item. Harnisch & Linn(1981) developed a modified version of C_i (called C_i^*) which has a lower bound of 0 and an upper bound of 1. Establishing the bounds about the index eliminates extreme scores which may be obtained on C_i.

The Modified Caution Index for the i^{th} examinee is:

$$C_i^* = \frac{\sum_{j=1}^{n_{i.}} (1 - u_{ij.})r_{ij.} - \sum_{j=n_{i.}+1}^{J} u_{ij.}r_{ij.}}{\sum_{j=1}^{n_{i.}} r_{ij.} - \sum_{j=J+1-n_{i.}}^{J} r_{ij.}} \qquad (2)$$

For the present study, a Modified Caution Index was computed for each individual using computer programs written by the author in the FORTRAN IV programming language. All statistical analyses were peformed using the Statistical Analysis System(SAS). An IBM 370/168 was used.

RESULTS

Tables 1(Reading) and 2(Mathematics) illustrate the mean Modified Caution Index(C_i^*) for students in each curricular program within each district for each of the three years. The first observation evident from the tables is that the mean indices for reading were larger than those for mathematics. Thus, there was a higher degree of unusual responses for the reading test than for the mathematics test. This result may be related to the greater complexity in teaching reading, especially reading comprehension, as compared to mathematics computation.

To examine the differences among the C_i^* for each situation, the students' reading and mathematics indices were used as dependent variables in partial hierarchical analyses of variance. The year tested and the students' district were crossed factors; the students' curricular program was nested within districts. Tables 3(Reading) and 4(Mathematics) present the results.

TABLE 1

MEAN MODIFIED CAUTION INDICES
FOR THE MBS READING TEST

District	1978					1980					1982				
	Business	Academic	Vocational	General	Total	Business	Academic	Vocational	General	Total	Business	Academic	Vocational	General	Total
Vocational A	.343 (5)	.242 (11)	.314 (180)	.307 (31)	.310 (227)	.317 (7)	.326 (3)	.304 (186)	.306 (17)	.304 (213)	.299 (12)	.338 (11)	.322 (188)	.302 (15)	.320 (226)
Rural B	.305 (21)	.341 (113)	.290 (7)	.291 (94)	.316 (235)	.317 (13)	.356 (114)	.360 (9)	.324 (101)	.340 (237)	.291 (7)	.332 (88)	.302 (7)	.344 (67)	.334 (169)
Suburban C	.303 (207)	.335 (226)	.310 (63)	.300 (107)	.315 (603)	.314 (172)	.321 (203)	.317 (81)	.306 (132)	.329 (588)	.317 (124)	.347 (199)	.341 (41)	.305 (110)	.329 (474)
Regional D	.343 (10)	.361 (254)	.315 (4)	.339 (65)	.355 (333)	.308 (8)	.381 (176)	.363 (7)	.353 (54)	.372 (245)	.337 (5)	.373 (152)	.431 (4)	.350 (77)	.366 (238)
Urban E	.294 (63)	.304 (220)	.313 (38)	.302 (128)	.303 (449)	.310 (88)	.317 (201)	.318 (44)	.318 (131)	.316 (464)	.287 (75)	.329 (178)	.290 (39)	.295 (121)	.308 (413)
Total	.304 (306)	.334 (824)	.312 (292)	.305 (425)	.319 (1847)	.313 (288)	.341 (697)	.312 (327)	.320 (435)	.325 (1747)	.306 (223)	.346 (628)	.321 (279)	.318 (390)	.328 (1520)

TABLE 2

MEAN MODIFIED CAUTION INDICES
FOR THE MBS MATHEMATICS TEST

District	1978					1980					1982				
	Business	Academic	Vocational	General	Total	Business	Academic	Vocational	General	Total	Business	Academic	Vocational	General	Total
Vocational A	.223 (5)	.258 (11)	.235 (180)	.237 (31)	.236 (277)	.252 (7)	.219 (3)	.264 (186)	.265 (17)	.263 (213)	.225 (12)	.315 (11)	.276 (188)	.268 (15)	.275 (226)
Rural B	.243 (21)	.301 (113)	.255 (7)	.262 (94)	.279 (235)	.263 (13)	.308 (114)	.278 (9)	.281 (101)	.293 (237)	.348 (7)	.366 (88)	.286 (7)	.343 (67)	.353 (169)
Suburban C	.222 (207)	.264 (226)	.239 (63)	.240 (107)	.243 (603)	.250 (172)	.271 (203)	.262 (81)	.254 (132)	.260 (588)	.277 (124)	.316 (199)	.298 (41)	.259 (110)	.291 (474)
Regional D	.278 (10)	.308 (254)	.305 (4)	.302 (65)	.306 (333)	.310 (8)	.321 (176)	.346 (7)	.287 (54)	.314 (245)	.254 (5)	.356 (152)	.304 (4)	.327 (77)	.343 (238)
Urban E	.243 (63)	.235 (220)	.250 (38)	.243 (128)	.240 (449)	.272 (88)	.264 (201)	.268 (44)	.242 (131)	.260 (464)	.261 (75)	.321 (178)	.251 (39)	.272 (121)	.289 (413)
Total	.230 (306)	.275 (824)	.239 (292)	.255 (425)	.257 (1847)	.259 (288)	.288 (697)	.266 (327)	.261 (435)	.272 (1747)	.271 (223)	.334 (628)	.276 (279)	.291 (390)	.303 (1520)

Table 3

Summary of the Analysis of Variance
For the Reading Modified Caution Index

Effect	D.F.	Sum of Squares	Mean Square	F
District	4	0.291	0.073	1.98
Program(District)	15	0.552	0.037	3.08*
Year	2	0.051	0.026	2.15
District*Year	8	0.116	0.015	1.22
Prog(Dist)*Year	30	0.355	0.012	1.04
Within Cell	5054	57.392	0.011	
Total	5113	60.285		

* p < .01

Table 4

Summary of the Analysis of Variance
For the Mathematics Modified Caution Index

Effect	D.F.	Sum of Squares	Mean Square	F
District	4	0.452	0.113	2.05
Program(District)	15	0.830	0.055	3.24*
Year	2	0.288	0.144	8.44*
District*Year	8	0.076	0.010	0.56
Prog(Dist)*Year	30	0.512	0.017	1.44
Within Cell	5054	59.648	0.012	
Total	5113	65.916		

* p < .01

There was no significant year effect for the reading test. However, there was such a significant effect for the mathematics test ($p < .01$). Scheffé's multiple comparison test showed that the mean C_i^* for the students tested in 1978 ($\bar{X} = 0.257$) was significantly smaller than that for 1980 ($\bar{X} = 0.272$) which was significantly smaller than the 1982 result ($\bar{X} = 0.303$).

Both 'year' results are fairly curious ones. A larger C_i^*

is associated with a more unusual response pattern, due in part perhaps to lack of curricular validity. One might reasonably expect that the C_i^*'s should decrease over time (i.e. as the skills are included in the curriculum) rather than either remain the same (reading) or increase (mathematics).

A possible explanation for these results is that since both tests' mean scores increased from 1978 to 1982, the increase was due to a better mastery of those skills included in the curriculum, but not of skills not in the curriculum. Thus, students were scoring higher in 1980 than they did in 1978 and higher in 1982 than they did in 1980. If higher scoring students missed easy items (which were not in their curriculum), their value of C_i^* would be greater than that for students with lower scores who missed the same items. This interpretation assumes that the curriculum did not change over time to reflect the tested material.

The significant curricular program effect ($p < .01$) for both reading and mathematics indicates that summed over the three years, there were significant differences in C_i^* for the various curricular programs within each district. This significant effect can be further analyzed using Scheffé comparisons. However that result would only identify the curricular program(s) which had significantly larger values of C_i^* than others. For purposes of examining curricular validity, it is more important to assume that the differences exist and to analyze the cause of the unusual response patterns, especially since the means were large.

A second series of analyses was conducted to identify the subsets of items which contributed most to the Modified Caution

Indices for each curricular programs and district for each year. Following the procedures of Harnisch & Linn(1981), the test results were evaluated using linear regression analyses. The proportion of students who correctly answered each item (p-value), was computed for each of the 110 reading and 95 mathematics items for each appropriate unit. Mean test performance is directly related to item p-values; thus, the regression analyses were performed on the p-values for each appropriate unit with the p-values from the state results for each year.

The expected item p-values for each unit were determined from the regression equation and a residual was computed for each item. Then items were categorized according to their content. The reading test was divided into its three clusters -- word recognition, reading comprehension and study skills; the mathematics test into its four -- computation, number concepts, measurement & geometry, and problem solving. Finer groupings of the items into the subskills which compose the clusters were not meaningful because each subskill is assessed by a very small number of items.

The mean residual for each cluster was computed and standardized by dividing it by the standard error of estimate. Those standardized mean residuals were multiplied by the square root of the number of items in the cluster. That resulted in weighted standardized mean residuals which, as Harnisch & Linn(1981), note are analogous to critical ratios. The weighted standardized mean residuals were used to compare the items in each cluster.

The first regression was performed on the p-values for each district. Table 5 reports those results for each district in each

of the three years. Values greater than 2.0 indicate that items in that cluster were much easier for the students in that school than would be expected from their overall performance and the relative difficulty of those items for the population of students in the particular year. A value less than -2.0 indicates that the items were much harder. Seven of the entries(6.7%) had weighted standardized mean residuals greater than 2.0 while 9(8.6%) had values less than -2.0.

TABLE 5

Weighted Standardized Mean Residuals Of District Item P-Values By Content Category For Each Year

District	Year	Content Category: Reading: Word Rec.	Reading: Read Comp.	Reading: Study Skills	Mathematics: Compu-tation	Mathematics: Number Conc.	Mathematics: Meas. .Geom.	Mathematics: Prob Solve
A	1978	1.86	-2.04	-0.11	1.24	-0.28	-1.37	-0.13
	1980	0.45	-0.33	0.06	2.54	0.43	-3.60	-0.35
	1982	1.58	-0.73	-0.69	-0.38	2.80	-0.88	-1.13
B	1978	-0.07	0.10	0.21	-0.06	-0.73	1.04	-0.35
	1980	-0.88	1.17	-1.14	-0.76	0.18	-0.28	1.32
	1982	0.77	-0.43	-0.19	-1.15	0.73	1.36	-0.35
C	1978	-3.13	1.33	1.62	0.09	-0.85	-1.81	2.70
	1980	-3.55	0.80	3.27	0.27	-0.73	-1.79	2.27
	1982	-2.01	1.22	0.32	-1.88	1.53	1.15	0.22
D	1978	-0.53	1.04	-1.63	-1.49	2.12	1.03	-0.89
	1980	0.52	0.66	-2.04	-3.44	0.56	2.53	1.99
	1982	-0.84	1.13	-1.13	-1.81	0.92	1.56	0.25
E	1978	1.55	-0.40	-1.23	0.07	0.03	-1.61	1.66
	1980	1.60	-1.78	1.36	0.59	-0.07	-2.11	1.51
	1982	-0.02	-0.39	0.80	1.26	1.15	-3.15	0.25

The interest lies with the large negative values. The most striking results from Table 5 are the residuals for which there were large negative values for all three years -- the Measurement & Geometry items for District A, the Word Recognition items for District C, The Study Skills and Computation items for District D, and the Measurement & Geometry items for district E. The consistently large negative entries for these areas were in contrast to the other districts' values for those clusters. Thus, these results suggest that the skills measured by those clusters may be included in the curriculum of the other districts, but not in the cited ones.

To further examine the results from Table 5, another regression analysis was conducted in which the unit of analysis was the curricular program within each district rather than the entire district. This analysis was conducted to examine whether there were differences in the mean residuals across the four types of programs. Table 6 presents these results for the districts and clusters which were noted as anomolous in Table 5.

As noted in Table 6, the large residuals persisted for District A's Measurement & Geometry items for all the curricular programs except for the College Preparatory one. Thus, it appears that Measurement & Geometry was emphasized more in the College Preparatory curriculum but not in the other three. District D's Study Skills items behaved in the same manner. Those skills may not have been stressed in the College Preparatory or General programs to the same extent that they were in the other two.

TABLE 6

Weighted Standardized Mean Residuals of Program P-Values For Certain Districts And Certain Content Categories

District Year	Instructional Program: Business/ Commercial	College Preparatory	Vocational	General
District A (Measurement & Geometry)				
1978	-1.77	-0.41	-1.43	0.02
1980	-0.70	2.86	-3.60	+1.39
1982	-0.87	0.80	-0.72	-1.79
District C (Word Recognition)				
1978	-2.49	-1.73	-1.64	-1.24
1980	-1.88	-3.81	-1.71	-1.04
1982	-1.34	-2.05	-0.68	-0.06
District D (Study Skills)				
1978	0.01	-1.30	-0.93	-1.90
1980	0.55	-2.34	0.62	-1.18
1982	-0.22	-0.84	-0.72	-1.33
District D (Computation)				
1978	-0.87	-1.69	-1.73	0.55
1980	-1.48	-3.47	-1.76	-1.35
1982	-0.52	-1.40	-0.70	-1.33
District E (Measurement & Geometry)				
1978	-2.61	0.42	-0.21	-1.92
1980	-2.64	-0.26	-2.12	-2.23
1982	-3.39	-1.15	-0.58	-3.13

A similar conclusion can be drawn for District E's Measurement & Geometry skills. The lower mean residual for the Business and General programs compared to the other two programs suggests a difference in the emphasis of these skills across programs. Finally, for District C's Word Recognition items and District D's Computation items, there was no discernible difference in the mean residuals across the programs, indicating no differences in the curriculum-to-test match across programs. However, the negative residuals did indicate a lower than expected performance.

It would be of interest to examine the relationships over time to determine the effect of the testing program's impact on changes in specific curricula. One can examine Tables 5 and 6 to determine trends of larger values of the residuals from 1978 to 1982. Yet, as previously stated, there is a confounding of increases in total test score which impacts on the values of the residuals. What one is able to conclude is that within each year the mean residuals reflects the relationship between that year's statewide performance and the expected performance of the units. Interpretations of comparisons among years may be tenuous.

Summary

This study examined the curricular validity of the New Jersey Minimum Basic Skills(MBS) test, a minimum competency test which is used for high school graduation decisions. Based on examinations of a Modified Caution Index, there were certain differences in the unusual response patterns. Further, the reading indices were larger than the mathematics indices, indicating that there may have been a greater match between the curriculum and the mathematics test than with the reading test.

There was also no difference in the mean Modified Caution Index for mathematics over time which could indicate a possible lack of improved consistency between the schools' curriculum and the content of the test. The C_i^* for reading increased significantly over time indicating perhaps a greater disparity between the curriculum and test -- certainly an anomolous and unexpected result given the importance placed on the MBS test by the public reporting of the results. The greater complexity in teaching

reading than mathematics as well as the improvement in scores from 1978 to 1982 are likely to be reasons for these results.

For both reading and mathematics, the mean Modified Caution Indices were reasonably large enough to suggest that there were very unusual response patterns. Regression analyses were conducted to examine the anomolous situations. The results of these analyses showed that there were differences between curricular programs within a school district in terms of the unusual response pattern. This result suggests that within districts there may be differences in the content coverage and emphasis placed on some of the subsets of items contained on an MCT.

It is not necessarily true that unusual response patterns are the result of a lack of a match between the content covered on a test and the curriculum. As Harnisch and Linn(1981) note there may be many explanations for the unusual patterns. Thus, one cannot conclude from this study that one Minimum Competency Test can(or cannot) be curricular valid for students in varying curricular programs in different districts. However, because differences were noted across districts and curricular programs, there is the suggestion that there may be problems using one test. Other, more detailed non-test based analyses should be conducted to further examine the curricular validity and also the instructional validity. Such information would be very beneficial for school districts to have for planning purposes.

The analyses conducted in this study provide an initial insight into possible differences between test and curricular matches. Such analyses are useful for detecting mismatches so

that corrective action can be taken. It students are to be held accountable by having their graduation decisions based in part on test results, it is critical that the test be content valid in its broadest sense.

REFERENCES

Guttman, L., The quantification of a class of attributes: A theory and method of scale construction. In P. Horst, P. Wallin, & L. Guttman (Eds.), The prediction of personal adjustment. New York: Social Science Research Council, Committee on Social Adjustment, 1941.

Haney, W., Validity and Competency Tests: The Debra P. case, conceptions of validity, and strategies for the future. In G. Madaus (Ed.), The Courts, Validity, and Minimum Competency Testing, Boston: Kluwer-Nijhoff Publishing, 1983.

Harnisch, D.L. & Linn, R.L., Analysis of item response patterns: Questionable test data and dissimilar curriculum practices., Journal of Educational Measurement, 1981, 18, 133-146.

Leinhardt, G., Overlap: testing whether it is taught. In G. Madaus (Ed.), The Courts, Validity, and Minimum Competency Testing, Boston: Kluwer-Nijhoff Publishing, 1983.

Madaus, G.F., Minimum competency testing for certification: The evolution and evaluation of test validity. In G. Madaus (Ed.), The Courts, Validity, and Minimum Competency Testing, Boston: Kluwer-Nijhoff Publishing, 1983.

New Jersey State Department of Education, New Jersey Minimum Basic Skills Testing Program, 1977-78: Directory of test specifications and Items, 1978.

New Jersey State Department of Education, New Jersey Minimum Basic Skills Testing Program, 1979-80: Directory of test specifications and Items, 1980.

New Jersey State Department of Education, New Jersey Minimum Basic Skills Testing Program, 1981-82: Directory of test specifications and Items, 1982.

New Jersey State Department of Education, New Jersey Minimum Basic Skills Testing Program, 1977-78: State report: Analysis and interpretation of statewide performance, 1978.

New Jersey State Department of Education, New Jersey Minimum Basic Skills Testing Program, 1979-80: State report: Analysis and interpretation of statewide performance, 1980.

New Jersey State Department of Education, New Jersey Minimum Basic Skills Testing Program, 1981-82: State report: Analysis and interpretation of statewide performance, 1982.

Pullin, D., Debra P. v. Turlington: Judicial standards for assessing the validity of minimum competency tests. In G. Madaus (Ed.), The Courts, Validity, and Minimum Competency Testing, Boston: Kluwer-Nijhoff Publishing, 1983.

Sato, T., [The construction and interpretation of S-P tables]. Tokyo: Meiji Tosho, 1975.

Schmidt, W.H., Porter, A.C., Schwille, J.R., Floden, R.E. and Freeman, D.J., Validity as a variable: Can the same certification test be valid for all students. In G. Madaus (Ed.), The Courts, Validity, and Minimum Competency Testing, Boston: Kluwer-Nijhoff Publishing, 1983.

Tatsuoka, M.M. Recent psychometric developments in Japan: Engineers grapple with educational measurement problems. Paper presented at the ONR Contractors Meeting on Individualized Measurement, Columbia, Mo., 1978.