ED 254 570 TM 850 187

AUTHOR Williams, Warren S.; Iverson, Bethany
TITLE Evaluating Locally D^eveloped Needs Assessment Measures.
PUB DATE Mar 85
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Admission Criteria; Cognitive Measurement; *Compensatory Education; Cutting Scores; Evaluation Methods; *High Risk Students; Primary Education; Psychomotor Skills; School Districts; *Screening Tests; Teacher Attitudes; *Test Reliability; *Test Validity
IDENTIFIERS *Taylor Public Schools

ABSTRACT

Four studies of the reliability and validity of needs assessment instruments developed by the Taylor Public Schools, Michigan, are described. The studies focused on the stability of student scores, classification stability, content validity, and concurrent validity. Consisting of separate tests for kindergarten, first and second grade, the instruments were designed to assist the school system in student selection for its compensatory education programs. Test scores were only modestly stable over a period of two to three weeks. The instruments did not reliably classify students for eligibility for compensatory education. Classroom teachers believed the tests measured important skills well. Correlations between teacher nomination for compensatory education and test scores were low. The studies indicate districts should place greater emphasis on teacher judgment rather than test scores to identify participants. Researchers expect that using cut off scores further from the median will result in improved classification stability. The reliability of selection test scores should be examined even if staff believes tests are adequately measuring important skills. The effectiveness of locally developed needs assessment instruments for student selection into compensatory education programs must be judged in comparison to alternative procedures. (DWH)

Evaluating Locally Developed Needs Assessment Measures

Warren S. Williams, Ed. D.
Eastern Michigan University

Bethany Iverson, M. A.
Taylor (MI) Public Schools

Introduction:

Local Educational Agencies (LEAs) typically use one of the following procedures to select children to participate in compensatory education programs:

1. They ask teachers to nominate children who need compensatory education services.

2. They use standardized tests to identify children performing below average compared to the national norms.

3. They use locally developed selection tests to identify children lacking basic skills.

While each of these approaches has strengths and weaknesses, many districts would prefer to use locally developed tests. These tests can be short, easy to administer and tied directly to the instructional needs defined by the district. However, locally developed needs assessment instruments have unknown psychometric properties, and districts are reluctant to base important selection decisions on these tests.

This paper describes four studies of the reliability and validity of needs assessment instruments developed by the Taylor (Michigan) Public Schools. Taylor is a predominantly working class suburb of Detroit and is the tenth largest school district in the state with approximately 13,500 students.

Needs Assessment Instruments:

The instruments consist of separate tests for kindergarten, first and second grade. (There is also a third grade test that was not included in this study.) Copies of the tests appear in Appendix A.

Each test consists of thirteen or fourteen items that measure if the child possesses the cognitive and psychomotor skills that teachers expect of students when a child enters a particular grade level. The tests are individually administered to all children at the beginning of each school year by experienced teachers.

The student's performance on each item is scored as either a "1" (100% accuracy), a "2" ("Some difficulty"), or a "3" (Poor performance). The number of 2's and 3's is used to determine if the child should participate in the compensatory education program. If the child has more than five or six scores of "2" or "3" (the threshold depends on the grade level), the child is considered for compensatory education services.

Brief Description of the Four Studies:

This paper summarizes the methodology and findings of four studies of the validity and reliability of these needs assessment instruments.

Study 1: Stability of Scores This study used a classical test-retest design to provide information about the stability of scores obtained on these tests.

Study 2: Classification Stability This study used the test-retest data to examine whether students maintained their classification of "needs compensatory education" or "does not need compensatory education" over a two week period.

Study 3: Content Validity This study examined the degree to which these tests measure concepts considered important by classroom teachers.

Study 4: Concurrent Validity This study examined the relationship between the scores of children on the needs assessment tests and teacher judgment about each child's need for compensatory education services.

The metric for analysis in these studies consists of the number of items on which the students did not perform satisfactorily. For example, a "score" of seven means the student failed seven items. Thus, the higher the students' scores, the pcorer their performance.

Confounding Variable:

The authors are aware of at least one important uncontrolled variable that might distort the results of this study.

Michigan regulations associated with its state-wide compensatory education program limit the number of children who can receive special services. The Taylor Schools, in an effort to operate the most effective program, adopted a student selection policy based on studies of the effect of early intervention on student performance. That is, the district believes it is important to identify children who are at academic risk while they are in the lower elementary grades; it concentrates its compensatory education effort on young children. Consequently, the district set cut-off scores on these tests at a level that would identify all students likely to have academic difficulty. This resulted in establishing low cut-off scores and in selecting some students who do not need compensatory education services. Teachers later had the opportunity to recommend that those children be dropped from the compensatory education program.

As one would expect, this early intervention policy caused the district to set cut-off scores that identified a disproportionate number of students as needing compensatory assistance. For example, 40% of the 965 students enrolled in kindergarten were selected for the compensatory education program. In first and second grades those percentages are 53% and 49% respectively.

The selection of this many children for the compensatory education program had two major effects on the present study. First, it forced the district to select children with relatively good scores to participate in the program. As one expects on mastery tests, there were many students who performed well on these measures, but the district set its cut-off scores within the midst of this large group. That had a detrimental impact on the reliability of the classifications of students into eligible and not eligible groups. Second, this decision lead to the selection of students that teachers might not ordinarily recommend for compensatory education services. This was particularly obvious in Study 4 in which teachers nominated far fewer children for compensatory education services than were admitted to the program based on their needs assessment scores. As mentioned earlier, the teachers later dropped students who should not participate in the compensatory education program.

Study 1: Stability of Scores

Test score reliability is a prerequisite for meaningful scores. This study used a classical test-retest design to examine the stability of the scores obtained by students on the three needs assessment measures.

Methodology:

All students in regular kindergarten, first and second grade classes took the needs assessment test in September, 1984. Twenty percent of the students who completed the pre-test were randomly selected from the total sample using a stratification procedure to insure a proportionate number of the students were selected from each school. Those students were re-tested between two and three weeks later by the same examiner. Pearson product-moment correlations were computed to estimate the test-retest reliability of the needs assessment scores.

Findings:

Table 1 presents the product-moment correlations of student test scores with scores obtained on the re-test. Correlations between .70 and .75 suggest that scores obtained on the tests are only modestly stable over a two to three week interval and that the district must be particularly cautious about using the results of this test to make decisions about the performance of individual students.

TABLE 1: TEST RE-TEST RELIABILITY

GRADE	N	r	SE_m
K	223	.73	1.7
1	228	.70	1.4
2	233	.75	1.4

There are at least two factors that contribute to these findings:

1. The tests are short, having between 13 and 14 items. Since it is unusual for a short test to generate reliable scores, the finding of modest score reliability should be expected for these needs assessment instruments.

2. Mastery tests of this type do not lend themselves to traditional correlational analysis (Gronlund, 1985). Examination of the score distributions indicate that most students performed well on the tests (as indicated by low mean scores on the two administrations of the test)[1]. In essence, there are two sets of scores. Most students performed well on the tests and their scores are clustered narrowly near the perfect score of zero. A smaller group did not perform well on the tests and their scores are distributed over a wide range. When these two sets of scores are combined, the result is a highly skewed distribution of scores that masks the restricted range of scores of the large group of successful students. This restriction in the range of scores of one group of students predictably reduces the correlation between the pre-test and post-test scores for the total group.

In summary, this test-retest reliability analysis suggests the tests yield results of low reliability. However, the procedures used in this traditional norm referenced approach to test reliability are possibly not appropriate for examining mastery tests.

Study 2: Classification Stability

The needs assessment tests developed by the district are selection measures; they classify students into two groups: students needing compensatory education assistance and those not needing assistance. Thus, these tests can be treated as mastery measures that students either pass or fail.

Gronlund (1985) suggests that psychometricians examine the stability of a mastery test by determining if the test is consistent in its ability to classify a student as passing or failing. For lack of a better term, we call this characteristic "classification stability"; the degree to which a test consistently classifies a student into the "needs help" and "does not need help" categories. If we assume that students who need compensatory education services at the end of September will also need those services two to three weeks later, reliable tests should consistently classify students as either needing help or not needing help over that time period.

Study 2 examined the classification stability of the needs assessment instruments.

[1] The reader is again reminded that low scores on the tests suggest high levels of student performance.

Methodology:

Data for this study,consists of a re-analysis of the data from Study 1, and subjects in this study are the same randomly selected 20% of the students attenaing kindergarten, first and second grade in the district. Each student was classified as "eligible" for compensatory education services and "not eligible" on both the test and the re-test administered two to three weeks later. Those data were the basis for the analysis.

Findings:

Tables 2-4 present the results of these analyses, and Table 5 summarizes the results of the studies at the three different grade levels. The metric labelled "Consistency" in Table 5 is the percentage of children classified into the same category ("eligible" or "not eligible") on both administrations of the test.

TABLE 2: CLASSIFICATION STABILITY - KINDERGARTEN

		ORIGINAL TEST		
		Eligible	Not Eligible	Total
RE-TEST	Eligible	60 68% 85%	11 8% 15%	71
	Not Elfgible	28 32% 18%	124 92% 82%	152
	Total	88	135	223

Cell Contents: N
Column %
Row %

Chi-square = 85.7

df = 1

$p < .01$

TABLE 3: CLASSIFICATION STABILITY - GRADE 1

		ORIGINAL TEST		
		Eligible	Not Eligible	Total
RE-TEST	Eligible	59 49% 83%	12 11% 17%	71
	Not Eligible	62 51% 39%	95 89% 61%	157
	Total	121	107	228

Cell Contents: N

Column %

Row %

Chi Square = 35.6

df = 1

p <.01

TABLE 4: CLASSIFICATION STABILITY - GRADE 2

		ORIGINAL TEST		
		Eligible	Not Eligible	Total
RE-TEST	Eligible	67 57% 83%	14 12% 17%	81
	Not Eligible	51 43% 34%	101 88% 66%	152
	Total	118	115	233

Cell Contents:
N
Column %
Row %

Chi Square = 49.1
df = 1
$p < .01$

TABLE 5: CLASSIFICATION STABILITY

% OF STUDENTS...

	Eligible On Test Who Were Eligible On Re-Test	Not Eligible On Test Who Were Not Eligible On Re-Test	Consistency[1]
K	68%	92%	83%
1	49%	89%	68%
2	57%	88%	72%

[1] Consistency = $\left(\frac{\text{Number of Students Whose Eligibility Didn't Change}}{\text{Total Number of Students}}\right) \times 100$

The results suggest that the tests have a modest ability to classify students into these categories with any degree of stability. Over all, approximately 25% of the students were classified into a different category when they took the same test two to three weeks later.

These results mask differences that exist between the students classified as "eligible" and those classified as "not eligible" based on the first administration of the test. As suggested by the data in Table 5, students classified as "not eligible" on the first administration of the test were far more likely to retain their "not eligible" classifications than were students classified as "eligible" on the first test.

Further examination of the data indicate that the re-test scores were substantially lower (i.e., better) than the scores obtained on the first administration of the measure. In kindergarten, 32% of the children who qualified for the program on the first administration of the test scored too low on the test to qualify for the program two to three weeks later. Perhaps this should be expected given the age of the children and the rate at which they learn introductory concepts. However, 51% of the first grade students who initially qualified for the compensatory education program did not qualify when given the same test two weeks later. In second grade, 43% of the students did not qualify for the program when re-tested.

Conversely, the results were stable for children who obtained good scores on the first administration of the test. Most of the children deemed not eligible for compensatory education on the first administration of the test were also not eligible on the second administration of the measure. Overall, only 10% of the children whose scores made them ineligible for the program on the original administration of the test were eligible on the re-test.

In general, one must conclude that while the tests provide reliable results for children who pass the test, the results are not sufficiently reliable for students who obtain poor scores on the measure. Since the purpose of the test is to identify children who quality for compensatory education services, and since 43% of those identified as "qualified" on the first administration of the test were "not qualified" two weeks later, we must conclude that the tests do not provide a consistent indicator of who should receive compensatory education services.

The authors have reservations about generalizing these findings to other compensatory education programs. As noted earlier, the Taylor Schools consciously over-identified children to participate in the compensatory education progwam at these lower grade levels. This led them to set cut-off scores that were close to the middle of the narrow cluster of scores obtained by students who performed well on the test. Thus the district selected and rejected many students who were close to the cut-off score; those students could change their position into "eligible" and "not eligible" groups based on test score changes of one or two points. It is likely that the district's decision had a detrimental impact on the tests' ability to reliably classify students into the two categories.[2]

[2] Readers might also suspect that part of the shift in test scores can be explained by statistical regression toward the mean. However, the decrease in number of students qualifying for the program is so dramatic that it is unlikely to be explainable by this phenomenon.

The authors are now conducting a study to examine the impact of using different cut-off scores on the classification stability of the needs assessment test results.

Study 3: Content Validity

The needs assessment tests were designed to measure whether students mastered the basic skills and concepts that are expected of children before they enter kindergarten, first and second grades. Study 3 examined (a) if teachers believe the skills and concepts measured on the needs assessment instruments are important for the educational development of children, and (b) the degree to which the locally developed needs assessment tests measured those concepts and skills.

Methodology:

Data for this study consists of teacher response to two questionnaires. One questionnaire measured teachers' perceptions of the importance of the skills tested on the needs assessment instruments; the second examined if teachers thought the tests measured those skills.

Items in the questionnaires were based on the skills being measured at each grade level. Since the intent of a test item is not always obvious from the item itself, the authors of the needs assessment instruments described the skill they were trying to measure in each item. Two questionnaires were developed for each grade level based on these lists of skills. One survey asked teachers to rate the extent to which each skill is important for students entering their grade level. The second survey presented a sample of each test item and a list of the related skills; the respondents indicated the degree to which each test item measured the related skill. Copies of these instruments appear in Appendix B.

All 83 kindergarten, first and second grade teachers in the district received a copy of the instruments. Seventy-ome teachers returned their questionnaires. Four of the questionnaires were discarded (The authors considered any questionnaire where all the ratings were identical as invalid. They also eliminated one questionnaire that had only four responses.), so 67 questionnaires were used in the final analysis, for an overall return ratio of 81%. Table 6 presents the number of questionnaires distributed and the return rate for each grade level.

TABLE 6: RESPONSE RATE FOR TEACHER QUESTIONNAIRES

GRADE	NUMBER DISTRIBUTED	NUMBER RETURNED	NUMBER CONSIDERED VALID	% OF VALID QUESTIONNAIRES
K	19	15	15	79%
1	31	29	26	84%
2	33	27	26	79%

Findings:

Tables 7 through 9 present the results of the two components of this study; Tables 10 and 11 summarize those results.

TABLE 7: CONTENT VALIDITY OF NEEDS ASSESSMENT MEASURES - KDG.

TEST ITEM	MEDIAN RATING	
	IMPORTANCE OF SKILL[1]	MEASURE OF SKILL[2]
#1 - Knows body parts	5	3
#2 - Fine motor skills - draw picture	4	4
#3 - Recognizes letters	4	2
#4 - Prints first name	4	5
#5 - Knows name	5	5
#6 - Recognizes letters	4	4
#7 - Knows abc's	5	5
#8 - Knows address	4	5
#9 - Answers with sentence	4	4
#10 - Repeats 4 words	4	5
#11 - Counts 1 to 10	4	5
#12 - Counts 4 objects	4	5
#13 - 1 to 1 correspondence	4	4
#14 - Color names	5	5
#15 - Knows body parts	4	5
#16 - Hops	4	4
#17 - Balances on 1 foot	4	4
#18 - Eye-hand coordination	4	5
#19 - Copies shapes	4	5

[1] Rating of Importance:
1 = Very unimportant
2 = Somewhat unim[illegible]t
3 = Neutral - nei[illegible] unimportant or important
4 = Somewhat important
5 = Very important

[2] Rating of Content Validity
1 = Definitely does not measure the concept
2 = Poor measure of the concep
3 = Neutral
4 = Good measure of the concep
5 = Excellent measure of the concept

TABLE 8: CONTENT VALIDITY OF NEEDS ASSESSMENT MEASURES - GRADE ONE

TEST ITEM	MEDIAN RATING	
	IMPORTANCE OF SKILL[1]	MEASURE OF SKILL[2]
#1 - Prints name	5	5
#2 - Recognizes letters	5	2
#3 - Knows body parts	4	3.5
#4 - Fine motor skills - draws picture	4	4
#5 - Knows complete address	4	5
#6 - Recites the alphabet	5	5
#7 - Knows upper case letters	5	5
#8 - Knows lower case letters	5	5
#9 - Remembers number sequence	4	5
#10 - Finds 3 matching letters	5	5
#11 - Counts to 20	5	5
#12 - Recognizes numbers 1 - 10	5	5
#13 - Selects 6 objects from 10	5	5
#14 - 1 to 1 correspondence	5	2
#15 - Knows geometric shapes	4	5
#16 - Knows missing numbers to 10	5	5
#17 - Skips	4	5

[1] Rating of Importance:
1 = Very unimportant
2 = Somewhat unimportant
3 = Neutral - neither unimportant or important
4 = Somewhat important
5 = Very important

[2] Rating of Content Validity:
1 = Definitely does not measure the concept
2 = Poor measure of the conc
3 = Neutral
4 = Good measure of the conc
5 = Excellent measure of the concept

TABLE 9: CONTENT VALIDITY OF NEEDS ASSESSMENT MEASURES - GRADE TWO

	MEDIAN RATING	
TEST ITEM	IMPORTANCE OF SKILL[1]	MEASURE OF SKILL[2]
#1 - Writes sentences	4	4
#2 - Prints letters	5	5
#3 - Knows body parts	5	4
#4 - Fine motor skills - draws picture	4	4
#5 - Knows missing numbers to 99	4	4
#6 - Adds numbers w/o regrouping	5	5
#7 - Subtracts w/o regrouping	5	5
#8 - Knows address, phone, birthday	5	5
#9 - Reads color words	5	5
#10 - Says sounds	5	5
#11 - Says short vowel sounds	4.5	5
#12 - Reads simple sentences	5	4
#13 - Reads analog time - hour	4	5
#14 - Reads numerals to 99	5	5
#15 - Solves addition word problems	4	5

[1] Rating of Importance:
1 = Very unimportant
2 = Somewhat unimportant
3 = Neutral - neither unimportant or important
4 = Somewhat important
5 = Very important

[2] Rating of Content Validity:
1 = Defiritely does not measure the concept
2 = Poor measure of the concept
3 = Neutral
4 = Good measure of the conc
5 = Excellent measure of the concept

TABLE 10: SUMMARY OF MEDIAN RATINGS ON A VALIDITY SURVEY - IMPORTANCE OF SKILLS

	Kdg.	Grade One	Grade Two
1. Very Unimportant	0[1]	0	0
2. Somewhat Important	0	0	0
3. Neutral	0	0	0
4. Somewhat Important	15	6	6
5. Very Important	4	11	9

[1] Interpretation: There were no skills measured on the Kindergarten test that received a median rating of 1 (very unimportant) on the teacher survey instrument.

TABLE 11: SJMMARY OF MEDIAN RATINGS ON VALIDITY SURVEY - ADEQUACY OF MEASUkE

	Kdg.	Grade One	Grade Two
1. Definitely doesn't measure concept	0[1]	0	0
2. Poor measure of concept	1	2	0
3. Neutral	1	1	0
4. Good measure of concept	6	1	5
5. Excellent measure of the concept	11	13	10

[1] Interpretation: There were no items on the Kindergarten test that had a medɸan rating of "1" from teachers who were asked to rate the test's ability to measurɾ the concept.

The data in Table 10 indicate that all 51 skills were rated "somewhat important" or "very important" by the teachers. Two thirds of the skills measured on the first and second grade tests were rated "very important" for incoming students at those grade levels. (The authors speculate that the differences between the kindergarten results and the findings for grades one and two reflects the lack of consensus that exists in the profession about the skills that should be brought to school by entering kindergarten students.) Overall, these data suggest that the tests measure skills considered important by teachers.

The data in Table 11 indicate the teachers believe the items were good to excellent measures of the skills they considered important for entering students. Sixty-seven percent of the items were rated as "excellent measures of the concept" and an additional 24% were rated as "good measures of the concept" by classroom teachers. Only 6% of the items were considered "poor measures of the concept".

The data summarized in Tables 7 through 11 suggest that the classroom teachers believe the tests do a good job of measuring concepts they consider important.

Study 4: Concurrent Validity

This study examined the relationship between each child's needs assessment score and teacher judgment about his/her need for compensatory education. It is based on the assumption that teachers can identify children with significant academic needs. If we accept that assumption, and if locally developed tests measure those needs, there should be a meaningful correlation between student's scores and the teacher's rating of their need for compensatory services.

Methodology:

Every kindergarten, first and second grade teacher in the district was asked to select those children in the class needing compensatory education services. These data and each child's test performance were used as the bases for two analyses:

1. A point bi-serial correlational analysis to examine the relationship between a dichotomous variable (the teacher's judgment of whether the child should be in the compensatory education program) and a continuous variable (the child's test score).

2. A chi-square analysis to determine if teacher judgments about who should receive compensatory education services correspond with the results of the needs assessment tests.

Findings--Correlational Analysis

Table 12 summarizes the correlations between teachers' nomination of students for compensatory education and student test scores on the needs assessment measures. While the p-values suggest these correlations are unlikely to occur by chance, the data indicate little relationship between teachers' perceptions of student needs for compensatory education and student scores on the locally developed needs assessment test .

TABLE 12 POINT BI-SERIAL CORRELATIONS BETWEEN TEACHER NOMINATIONS AND STUDENT SCORES

Grade	N	r	df	p
K	858	.42	856	$<.01$
1	884	.42	882	$<.01$
2	894	.50	892	$<.01$

Findings--Chi-square Analysis

Tables 13-15 present cross tabulations comparing teacher judgment of student needs for compensatory education and the selection of students into the program based on the needs assessment instruments. Each child was classified into one of four cells. For example, the upper left-hand cell indicates the number of students who were eligible for compensatory education based on the needs assessment test results and were also recommended for compensatory education by their teacher. Each cell contains: (a) the number of students in that cell, (b) the number of students "expected" in that cell (using the standard chi-square technique to generate expected cell frequencies) and, (c) the proportion of students from that column in the cell. The last figure reflects the degree of agreement between the test results and teacher judgment.

TABLE 13: COMPARISON BETWEEN TEST RESULTS AND TEACHER NOMINATIONS FOR COMPENSATORY EDUCATION - KINDERGARTEN

		TEST RESULTS		
		Eligible	Not Eligible	Total
TEACHER NOMINATION	Nominated	150 (83) 45%	66 (133) 13%	216
	Not Nominated	180 (247) 55%	461 (394) 87%	641
	Total	330	527	857

Cell Contents:

N (Expected N) Column %

Chi-Square = 115

df = 1

p .01

TABLE 14: COMPARISON BETWEEN TEST RESULTS AND TEACHER NOMINATIONS FOR COMPENSATORY EDUCATION - GRADE 1

		TEST RESULTS		
		Eligible	Not Eligible	Total
TEACHER NOMINATION	Nominated	234 (161) 50%	70 (143) 17%	304
	Not Nominated	235 (308) 50%	345 (272) 83%	580
	Total	469	415	884

Cell Contents:

N

(Expected N)

Column %

Chi-Square = 105

df = 1

p <.01

TABLE 15: COMPARISON BETWEEN TEST RESULTS AND TEACHER NOMINATIONS FOR COMPENSATORY EDUCATION - GRADE 2

TEST RESULTS

TEACHER NOMINATION		Eligible	Not Eligible	Total
	Nominated	240 (153) 54%	70 (157) 15%	310
	Not Nominated	201 (288) 46%	383 (296) 85%	584
	Total	441	453	894

Cell Contents:

N
(Expected N
Column %

Chi-Square = 148

df = 1

p <.01

All three analyses yield chi-squares associated with p-values of <.01. Consequently, we must reject the null hypothesis and conclude that the differences between these two bases for selection into the compensatory education program are unlikely to be attributable to chance.

If we accept the assumption that teachers can identify children who need compensatory education services, these data support three conclusions:

1. There is a meaningful difference between teacher judgment about who should be in the program and the students who were actually selected for the program using the needs assessment test results. Overall, approximately 50% of the students selected to participate in the compensatory education program based on test results were not nominated for the program by their teachers.

2. Teachers nominated fewer children for the compensatory education program than were selected by the needs assessment tests. For example, Table 14 shows that while 469 students were selected into the first grade compensatory education program based on their needs assessment tests, teachers recommended only 304 children for that program.

3. Teachers generally concurred with the results of the needs assessment tests for students who were not eligible for the program. Overall, the teachers indicated that approximately 85% of the children excluded from the program based on their needs assessment test scores should not be in the program. However, a small but meaningful number of students who were nominated for the program by their teachers did not have test scores that qualified them for the program.

In part, these results reflect the policy decision by the Taylor Schools to provide compensatory education services to a large group of students in the early elementary grades. As described earlier, that decision led to (a) the acceptance of significantly more students into the program than would be nominated by classroom teachers, and (b) use of a cut-off scores that were closer to the average score. Use of the lower (i.e., less restrictive) cut-off score led to many students being admitted or rejected from the program based on a one or two point difference from the cut-off score. This suggests that the findings of this study might be different if the district selected a cut-off score only slightly higher or lower than the one actually used. The researchers are presently conducting a series of studies on the impact of using different cut-off scores on the reliability and concurrent validity of the needs assessment instruments.

Summary:

This paper describes four studies of some of the psychometric properties of short, locally developed needs assessment instruments. These tests were designed to help the Taylor Public Schools select students for its compensatory education programs. The studies, which focused on the stability of student scores, classification stability, content validity and concurrent validity, support the following conclusions:

1. The test scores are only modestly stable over a period of two to three weeks.

2. The tests did not reliably classify students into the "eligible" for compensatory education and "not eligible" groups. Approximately 43% of the students classified as "eligible" on the first administration of the test were "not eligible" when re-tested on the same measure two to three weeks later.

3. Classroom teachers believe the tests do a good job of measuring skills they consider important for incoming students. The teachers rated all the skills tested by these items as "somewhat important" or "important", and 91% of the items were considered "good" or "excellent" measures of those concepts.

4. There is only a modest relationship between a teacher's belief that a student should participate in the compensatory education program and the test score obtained by that student. The correlations between teacher nomination and test scores is low and approximately 50% of the students selected for the program based on the needs assessment test scores were not nominated for the program by their teachers. The researchers describe how using inappropriate cut-off scores might affect the findings of these studies.

Conclusions:

This work raises several issues that merit further examination. First, the researchers found significant inconsistencies between teacher judgments about student need for compensatory education and the results of the district's needs assessment instruments. If we assume that teacher judgment is a reliable and valid indicator of student need, then these findings question the efficacy of the locally developed needs assessment instruments. But if teacher judgment is unstable or invalid, we cannot conclude that the needs assessment instruments are faulty. Certainly the issue of teacher nomination for compensatory education needs further study. If teachers generate appropriate lists of children to receive compensatory education services, perhaps districts should place greater emphasis on teacher judgment instead of emphasizing the use of tests to identify participants.

Second is the issue of the impact of using particular cut-off scores on the reliability of selection instruments. The researchers are reanalyzing the Taylor data to determine the impact of using different cut-off scores on the classification stability of the instruments. They expect that using cut-off scores further from the median will result in improved classification stability.

Third is the disparity between teacher judgment about the quality of the items on the tests and the generally unreliable results yielded by the measures using the present cut-off scores. Lack of test reliability sets a cap on the validity of scores generated by a measure. The fact that teachers believe the tests do a good job of measuring important skills should be viewed with caution if, in fact, a test yields generally unstable results. These findings suggest that districts examine the reliability of their selection test scores, even if the staff believes the measures are doing a good job of measuring important skills. However, it should be recognized that the judgment of these teachers might be correct; the tests might yield valid results if more appropriate cut-off scores are established.

Finally, there is the issue of how to best select children for compensatory education programs. As suggested earlier, using locally developed needs assessment instruments is one of several alternative procedures. The effectiveness of these i struments should be judged in comparison to the alternatives. For example, norm referenced achievement tests and teacher rating instruments are widely used for student selection into compensatory education programs. Additional work should be done to study the classification stability of these measures when they are used as a basis for student selection.