DOCUMENT RESUME

ED 228 328 TM 830 266

AUTHOR Jones, Eric D.; And Others
TITLE Out-of-Level Testing for Special Education Students with Mild Learning Handicaps.
PUB DATE Apr 83
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Diagnosis; *Error of Measurement; Guessing (Tests); Intermediate Grades; Learning Disabilities; *Mild Disabilities; Reading Achievement; Scaling; Scoring; *Special Education; Test Reliability; *Test Use; Test Validity
IDENTIFIERS California Achievement Tests; Chance Level Scores; *Out of Level Testing

ABSTRACT

The purpose of this study was to evaluate the utility of out-of-level testing (OLT) when it is applied to the assessment of special education students with mild learning handicaps. This evaluation of OLT involved testing hypotheses related to: (1) the adequacy of vertical scaling, (2) the reliability and (3) the validity of OLT scores. Fifty-eight special education students were tested. All students had measured reading achievement below the 20th percentile, and were integrated into the regular fifth or sixth grade classrooms for a portion of their academic programs. The appropriate in-level tests and two consecutively lower OLTs of the reading subtest of the California Achievement Test were administered approximately 10 days apart. The results suggest moderate support for the utility of OLT. It is suggested that the congruence between test content and the instructional programs may have considerably more influence over the reliability and validity of test data than would the use of OLT. (Author)

OUT-OF-LEVEL TESTING FOR SPECIAL EDUCATION STUDENTS WITH MILD LEARNING HANDICAPS

Eric D. Jones

Bowling Green State University

Bowling Green, Ohio

J. Jackson Barnette and

Carolyn M. Callahan

University of Virginia

Presented at the Meeting of the American Educational Research Association, Montreal, 1983

*Printed in U.S.A.

Abstract

The purpose of this study was to evaluate the utility of out-of-level testing (OLT) when it is applied to the assessment of special education students with mild learning handicaps. This evaluation of OLT involved testing hypotheses related to (a) the adequacy of vertical scaling, (b) the reliability and (c) the validity of OLT scores. Fifty eight special education students were tested. All students had measured reading achievement below the 20th percentile, and were integrated into the regular fifth or sixth grade classrooms for a portion of their academic programs. The appropriate in-level tests and two consecutively lower OLTs of the reading subtest of the California Achievement Test they were administered approximately 10 days apart. The results suggest moderate support for the utility of OLT. It is suggested that the congruence between test content and the instructional programs may have considerably more influence over the reliability and validity of test data than would the use of OLT.

OUT-OF-LEVEL TESTING FOR SPECIAL EDUCATION STUDENTS WITH MILD LEARNING HANDICAPS

There is considerable overlap between Title I and special education programs with respect to: populations served, instructional approaches, and program goals (Birman, 1981). There is not a substantial overlap of evaluation procedures. The basis of Title I program evaluation is group-level norm referenced achievement test data. Special education achievement evaluation emphasizes individual testing. Although group-level evaluation has obvious utility in special education program evaluation, very little attention has been given to the improvement of procedures for group-level evaluation of special education programs.

Out-of-level testing (OLT) is a norm referenced testing procedure which was developed for Title I evaluation. It is intended to allow low achieving students to be evaluated with test levels which more closely match their skills than would the test levels recommended for their grade level peers. Out-of-level testing of low achieving students is possible because most major publishers of achievement tests develop a series of tests which are organized into levels. Each test level is designed to test a domain of skills appropriate to a particular range of grade levels. Frequently, the level recommended by the publisher is too difficult for extremely low achieving students. In such cases out-of-level testing would allow for testing low achieving students on a lower level test. A vertical equating of scores across test levels allows norm referenced comparisons between students who took different test levels. The purposes for using out-of-level testing is to improve the reliability and validity of norm referenced scores of low achieving students.

Despite its wide use in Title I program evaluation, OLT has not received very much use in special education program evaluation. There are two reports of the use of OLT for the evaluation of mainstreamed special education students (Meyers, MacMillan, and Yoshida, 1975; Yoshida, 1976). The procedures reported by Meyers et al. and Yoshida are not comparable to those used in the evaluation of Title I programs. First, subjects were assigned test levels which were frequently two or more levels below the age-grade level test. Among Title I evaluators, it is generally recognized that testing more than one level out-of-level is difficult to justify. Some tests such as the California Achievement Test may be used to test one or two levels below the level recommended for in-level testing. With such tests, each test level is intended to assess performance within only one grade. Tests which are divided in levels which are each appropriate for the in-level evaluation of two or more grades are less apt to be valid when used to test more than one level below the recommended in-level test. Vertical equating is more likely to be better and content across levels is apt to be more similar for tests which have several levels which evaluate a relatively narrow range of grades than for tests which evaluate a broader range of grades. The second shortcoming of those studies was a failure to convert obtained scores into in-level percentiles which made relative comparisons impossible. Third, no attempt was made to assess the adequacy of the vertical equating of test levels. The results of studies by Meyers et al. (1975) and Yoshida (1976) failed to justify the use of out-of-level testing procedures for populations with mild learning handicaps.

Studies of OLT with Title I samples have not uniformly supported the logic of OLT. Those studies have usually focused upon reductions in the

proportions of students scoring below the chance-level or upon the adequacy of vertical scaling (Arter, 1980). A problem with the interpretation of those studies has been a lack of agreement on the criteria for evaluating the adequacy of vertical scaling. Other limitations to interpretation may be related to various methodological weaknesses. Despite methodological weaknesses and conflicting interpretations, the Title I studies have indicated some support for the adequacy of out-of-level procedures. That support would not warrant generalizations of results from Title I studies to special education studies.

The purpose of this study was to evaluate the utility of out-of-level testing when it is applied to the assessment of special education students with mild learning handicaps. The evaluation of OLT when applied to special education evaluation involved testing hypotheses related to the adequacy of vertical scaling, the reliability and the validity of out-of-level scores. The seven research hypotheses for this study are listed below.

1. The mean vertical scale scores will be significantly lower for out-of-level tests (OLT) than for in-level tests (ILT).
2. The Kuder-Richardson 20 estimates of internal consistency will be substantially greater for out-of-level than for in-level tests.
3. The proportions of raw scores above the chance-level will be significantly higher for the out-of-level tests than for the in-level tests.

4. The proportion items which tended to receive random responses will be significantly greater for the in-level than for the out-of-level tests.
5. The comparison of median in-level and out-of-level point-biserial correlations (pbis r) will indicate that items on the out-of-level test offer better predictions of high and low scoring subjects.
6. The correlations between vertical scale scores and teacher estimates of current reading instructional levels will be higher for the out-of-level tests than for the in-level tests.
7. The regression of the in-level test on the out-of-level test will indicate very low predictive value for the in-level test.

Method

Sample

Fifty-eight students from fifth and sixth grade resource rooms and part-time special education classrooms participated in the study. The sample was non-random--that is, all available subjects were used. The sample was obtained from 11 rural and suburban elementary and middle schools. All of the subjects in the sample met the following criteria: (1) placement was in an integrated special education program (educational services were provided in both the regular and special education classrooms), (2) measured reading achievement was below the 20th percentile or two years below grade level, (3) regular classroom services were provided for a portion of the educational program, and (4) all subjects were males. A summary of the demographic data is presented in Table 1. Prior to testing, three students who otherwise met the above criteria were eliminated from the study. Reasons for excluding them from

the study included: prominent behavior disorders, use of medications to control hyperactive behavior or epileptic seizures, and salient perceptual disorders or sensory impairments.

- - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - -

Most of the students who met the above criteria were classified as learning disabled (LD), 91.3 percent. The remainder of the students included 6.9 percent emotionally disturbed (ED) and 1.8 percent educable mentally retarded (EMR). As long as the special education students met the previous criteria, their classification of ED, EMR, or LD was not an overriding concern. Hallahan and Kauffman (1977) discussed the overlaps in definition, etiology, behavioral characteristics and instructional methods among ED, EMR, and LD students. They concluded that the present classification system does not provide for succinct classifications of students. The differences between the ED, EMR, and LD classifications appear to be quantitative with considerable overlap. Generally, regardless of classification students would best be served by the same general instructional approaches. Samples were drawn across school districts, and it appeared that differences in student classifications were more apt to be related to the particular LEA than to differences in the behaviors of mildly handicapped students.

Measures

Achievement measures. The reading subtest of the California Achievement Test (CAT, 1978) was selected for use in this study. For 6th graders the in-level test (ILT) was level 16 of the CAT. Level 15 was the

first out-of-level test (OLT1) and level 14 was the second out-of-level test (OLT2) for the 6th graders. For 5th graders the ILT was level 15. The OLT1 as level 14 and the OLT2 was level 13 of the CAT. Only one student of the 58 missed a test level (OLT2) due to absenteeism. Tests were administered approximately ten days apart during the spring of 1982.

The CAT was chosen for a number of reasons. First, it is one of the most widely used norm referenced tests in both regular educational testing and Title I evaluation. The favorable reviews of the 1970 version of the CAT in Buros (1978) would suggest the popularity of the 1978 version is justified.

Second, the publisher gave special consideration to developing the test with the intent that it be used for out-f-level testing. For example, locator tests were developed to aid in the selection of the appropriate test level for each student. The publisher also provided that students may be tested as far as two levels out-of-level with the CAT. That additional flexibility may be important in applications to special education populations where achievement would be expected to be lower than in Title I populations.

The third consideration was related to the outcome of out-of-level testing studies which used the CAT. Stewart (1980a) found no significant inconsistencies between in- and out-of-level testing with the CAT. That finding may have indicated that either there were no serious violations of the assumptions of the Thurstone scaling procedure, or that characteristics of the sample tested by Stewart (1980a) concealed any problems with the original scaling. That distinction cannot be made, but the suggestion that the vertical scaling of the CAT was adequate appears plausible.

The fourth consideration which supported the choice of the CAT was that during the standardization, students who were identified as special education students, but attended regular classes were tested (Technical bulletin 2, 1980). The extent to which the inclusion of such special education students is significant would be hard to evaluate. CTB/McGraw-Hill, the publisher of the CAT, did not develop separate norms for special education students. Other major publishers have also included special education students with mild learning handicaps in the revised standardizations of their achievement tests, but separate norms were not developed for those tests, either. In summary, the CAT was selected on the basis of its popularity, favorable expert opinions, technical features, and rather favorable research findings.

Teacher estimates of achievement. The special education teachers estimated the reading levels of the students they were responsible for teaching. Estimates were based upon the levels of the students' curricular materials and the students' proficiency with those materials. Estimates were scaled as grade equivalents. The estimated reading levels are presented in Table 1.

Several actions were taken to avoid potential sources of bias. First, to avoid bias due to students being aware of whether or not they were taking the most difficult test or one of less difficulty, a gummed label with the student's name or identification number was placed over the level number on the test booklet. Second, to avoid clerical errors, machine scorable answer sheets were not used. All answers were marked directly in the test booklets. Third, to avoid an order effect, the orders in which students took the different levels were counterbalanced. There were six different testing orders, and subjects were randomly

assigned. Fourth, students were tested in small groups of two to five per group. The testing rooms were quiet and work space was adequate. The students were seated so that copying, obvious guessing, and overt off-task behaviors could be prevented.

Analyses

Eight main analyses were performed for this study. Before testing any of the hypotheses, a preliminary analysis (a one-way ANOVA) was conducted to determine whether or not the mean raw scores of the in-level tests differed significantly between six different orders for test administration. A significant difference between the means would indicate that the results would be confounded by the orders of test administration. Because both tests had possible raw score totals of 70 points, grades five and six were combined for this analysis.

A two-way ANOVA with repeated measures was used to test the significance of differences between the mean vertical scale scores of the in- and out-of-level tests. The Scheffe test was used as the follow-up analysis.

Three analyses were conducted to test hypotheses related to the reliability of out-of-level test scores. First, KR-20 reliability coefficients were computed in order to estimate and compare the internal consistencies of the in- and out-of-level tests. Second, the Cochran Q Test was used to determine whether the proportions of scores below the chance-level (25 percent correct) decreased significantly with out-of-level testing. The McNemar test was used to make the follow-up comparisons. Third, an assumption was made in order to determine the proportion of items on which subjects tended to guess at random. The assumption was that if the proportion of responses to the alternative

answers of a multiple choice question did not vary significantly, then the item was one at which the students as a group tended to guess. To test whether or not students guessed at items, a chi-square test of the homogeneity of proportions was calculated on each item for each test level in both grades. The proportions of guessed items were tabulated for both test levels and grades. The K-sample Binomial Test of equal proportions was used to determine whether or not the frequencies of guessing differed significantly for in- and out-of-level tests. The McNemar Test was used for the follow-up comparisons of differences between test pairs.

Three analyses were conducted to test the hypotheses related to the validity of out-of-level test scores. First, in order to make estimates of item validity, the point-biserial item to total test correlation coefficients (pbis r) were calculated for each test level at each grade. The medians for the pbis r distributions were then calculated and compared to the medians obtained during the national norming of the CAT. Second, the Spearman rank correlation coefficient was used to determine the strength of the relationship between teachers' estimates of student achievement in reading (expressed in grade equivalents, GE's) and the vertical scale scores obtained by students on the in- and out-of-level tests. Third, at each grade level the out-of-level tests were regressed on the in-level tests to determine if the vertical scale scores obtained on the out-of-level tests could be predicted by the in-level vertical scale scores. These regressions were based upon combined grades.

Results and Interpretations

The results of the one way ANOVA to test for an effect of test order revealed that there was not a significant difference between the in-level raw scores across the six different orders of test administration, F (df 5,52) = .0818, p = .9948.

Adequacy of Vertical Scaling

The most popular approach to the assessment of vertical scaling has been to test the significance of differences between the in- and out-of-level mean vertical scale scores. In this study, not only were the differences between means considered, the standard errors of measurement associated with the means of each test level were also considered. The two-way ANOVA with repeated measures revealed statistically significant main effects for grades and for test levels (see Table 2). The mean vertical scale scores for the sixth graders were higher than for the fifth graders. The decline in vertical scale scores with out-of-level testing was also statistically significant. In the follow-up analyses the Scheffe test was used. The only significant difference between means for test levels was between the ILT and OLT2 pair, F (df 2,55) = 3.19, $p < .05$. The differences between other paired contrasts of adjacent test levels were not significant at the $p < .05$ level of significance. These results were consistent with those of earlier studies (cf. Crowder, 1978; Pelavin & Barker, 1976; Slaughter & Galles, 1978). It appears that the magnitude and direction of change in vertical scale scores depend in part upon the students' levels of achievement and the tests they took. It was assumed that if the floor of a test were removed, then the mean vertical scale scores would decrease significantly. If that assumption is reasonable, then it appears that for low achieving students in this study,

the floor effect may not begin to disappear until the students were tested two levels below the in-level test.

- - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - -

The comparisons of the standard errors of measurement associated with vertical scale score means indicated reductions in the amount of error associated with the measurement at each level (see Table 3). Those results also suggested that the differences between means, which were identified by the ANOVA procedure, would be of little practical importance. The rather large amounts of error associated with the means at each level threatened the validity of concluding that statistically significant differences between mean vertical scale scores existed at any of the levels.

- - - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - - -

These results do not clearly demonstrate the adequacy of the vertical scaling nor do they support an argument against it. Although researchers have not agreed as to whether or not mean vertical scale scores should drop or remain the same with out-of-level testing, they are in general agreement that the scale scores should not rise. The differences between the mean vertical scale scores and the standard errors of measurement were all consistent with the predicted direction of changes, and it does appear that there is a significant decrease in the amount of error from the in-level to the out-of-level tests.

Reliability

The KR-20 reliability coefficients are presented in Table 4. Two observations can be made based upon those data. The first observation is that the coefficients are all fairly high. The reliability coefficients for the in-level tests are sufficiently high to evaluate the level of large group achievement (.50 or greater), but fall far short of the magnitude that would be needed to make evaluations of individual achievement (Kelly, 1927; cited in Helmstadter, 1964). The second observation is that while the changes are in the predicted direction, they are not large. If in-level test reliabilities are sufficiently high and similar to out-of-level reliabilities, then a rationale for testing out-of-level must be based upon considerations other than a general desire to increase the reliabilities of achievement measures. Such considerations would be the achievement level of the group being tested and the content being considered. With this sample, the raw scores were higher than expected. Twenty-one percent of the fifth graders and 26 percent of the sixth graders got 45-75 percent of the items correct on the in-level test. A lower achieving group might have had lower internal consistency reliabilities for the in-level test, and would, perhaps, benefit more from assessment with an out-of-level test than did the group sampled in this study. It would also be worthwhile to consider the content of the out-of-level test. Even though out-of-level testing might yield an increase in reliability, it might do so at the expense of content validity.

- - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - -

In previous studies inferences about improved reliability with OLT were made by comparing the proportions of chance-level scores between test levels (e.g., Ayer & McNamara, 1973; Barnes, 1977; Mendro, 1977; Slaughter & Galles, 1978). The validity of such comparisons depends upon the validity of the assumption that chance-level scores are apt to be obtained by guessing. There is some question as to the plausibility of that assumption (cf. Stewart, 1980b). Comparisons of the frequencies of scores below 25 percent correct are presented in Table 5. The results of the Cochran Q Test for significance of changes in the proportions of raw scores below 25 percent correct are presented in Table 6. Results of the omnibus tests indicate that the proportions of raw scores below the chance-level were significantly lower for the out-of-level tests than for the in-level tests with fifth grade students, but not with sixth graders.

- - - - - - - - - - - - - - - - - -

Insert Tables 5 and 6 about here

- - - - - - - - - - - - - - - - - -

The McNemar Test was used as the follow-up test to make paired-comparisons of the increased proportions of scores above the chance-level. The data in Table 7 reveal that in the fifth grade there was a significant decrease in the proportions of scores below the chance-level between the ILT-OLT2 pair and the OLT1-OLT2 pair. Two

- - - - - - - - - - - - -

Insert Table 7 about here

- - - - - - - - - - - - -

observations can be made based upon those results. First, the proportions of students who scored below the chance-level on the in-level tests were not very large--21.4 percent of the fifth grade and 13.3 percent of the sixth grade. Integrated special education students with mild learning handicaps are generally considered to be lower achievers than their peers in Title I programs (cf. Birman, 1981). Therefore, higher proportions of chance-level scoring were expected than were actually observed. Second, although the proportions of chance-level scores decreased with out-of-level testing, the decreases were only significant between the ILT-OLT2 and OLT1-OLT2 test pairs for the fifth graders. There was not a significant reduction in chance-level scores for the sixth graders. The results of analyses of chance-level scores do not offer dramatic support for the value of out-of-level testing in reducing chance-level scores. The results of the McNemar analyses revealed that the changes in the proportions of scores below the chance-level were not statistically significant. These results are consistent with those of Mendro (1977) and Slaughter and Galles (1978), but perhaps the results are not necessarily supportive of those studies. As indicated in Table 5, the number of scores below the chance-level is quite small.

The validity of comparisons of the proportions of chance level scores with in- and out-of-level testing depends upon the assumption that chance level scores are apt to be obtained by guessing. Stewart (1980b) questioned the plausibility of that assumption. He argued that the notion of chance-level scores was based upon an inappropriate model of examinee

behavior. Stewart did suggest that it is plausible that as a group students may tend to guess at particular items. Such items would be identified by analyzing the responses to each item using a chi-square test of independence. Items with non-significant chi-square values could be considered to be items which are frequently guessed.

The frequencies and proportions of items with non-significant chi-square values for alternative responses are presented in Table 8. The results of the K-sample Binomial Tests for Equal Proportions revealed statistically significant decreases in the proportions of guessed items with out-of-level testing for both grades [fifth grade: chi-square (df 2) = 36.6, $p < .05$; sixth grade: chi-square (df 2) = 6.873, $p < .05$]. The paired comparison follow-up tests (see Table 9) reveal that there were significantly fewer guessed items for fifth graders on the OLT2 test than on either the ILT or the OLT1 test. For sixth graders the proportion of guessed items was significantly lower for the OLT2 test than for the ILT. That difference was the only one that was significant for the sixth

- - - - - - - - - - - - - - - - - -

Insert Tables 8 and 9 about here

- - - - - - - - - - - - - - - - - -

graders. The results of comparisons of the proportions of guessed items offer more support for the value of out-of-level testing in reducing guessing than is offered by the comparisons of chance-level scores.

The results of the two approaches to the analysis of guessing do not conflict with each other. It appears that the procedure recommended by Stewart (1980b) is more sensitive to the effects of out-of-level testing on guessing than the comparisons of chance-level scores. Since neither approach revealed a significant reduction for the ILT-OLT1 pair, it may be

concluded that testing only one level below the in-level test is unlikely to reveal either a reduction in the number of extremely low scores or a reduction in the number of items at which the students appear to guess.

Validity

An extensive investigation of the validity of out-of-level testing was beyond the scope of this study. The treatment of validity was limited to item-validity and two aspects of predictive validity. Item-validity was assessed by comparisons of the median pbis r's and their confidence intervals. Concurrent validity was examined through (a) the comparison of the Spearman-rank correlations of teacher estimates of current instructional levels with the obtained vertical scale scores for each test level and (b) the prediction of out-of-level test scores from the in-level test scores.

The comparisons of median pbis r's suggested that there was not a dramatic increase in the item-validity at each level (see Table 10). It may be that the difference between pbis r's of about .50 and pbis r's of .40 or lower would be of some practical significance. Henrysson (1971) described tests with average pbis r's of .40 or less as being very heterogeneous with respect to the traits being measured. Tests with pbis r's of .50 were regarded as moderately heterogeneous. It appears, however, that the changes in the pbis r's are not of statistical or practical significance. Yoshida (1976) reported what might have been considerably higher pbis r's for out-of-level testing with the 1970 version of the Metropolitan Achievement Test, but the format of his data tables did not allow for adequate description of the distributions of his pbis r's. Furthermore, students in his study were tested two or more levels out-of-level. Based upon the results of this study, it does not

appear that there is substantial improvement in item-validity as a result of out-of-level testing.

- - - - - - - - - - - - - - -

Insert Table 10 about here

- - - - - - - - - - - - - - -

The comparison of the Spearman rank correlations indicates that the relationships between teacher estimates of current instructional levels and the vertical scale scores earned on each level were quite similar (see Table 11). Since it was not possible to estimate the amount of error in the teacher estimation, that analysis is not a particularly rigorous assessment of predictive validity. The correlations may, however, offer an indication of the degree to which teachers would be apt to consider the tests as providing accurate evaluations of the students' achievement. Since one of the presumed benefits of out-of-level testing is that the evaluation process is more acceptable for both low achieving students and their teachers, information on the change in teacher acceptance would be useful. If the correlations between teacher estimates of achievement and earned test scores suggest that kind of information, they also suggest that teacher acceptance of in- and out-of-level test scores would probably not differ appreciably. Based upon the results of this study, one would not predict that teachers will be anymore favorably disposed toward out-of-level test scores, than they are to in-level scores.

- - - - - - - - - - - - - - -

Insert Table 11 about here

- - - - - - - - - - - - - - -

The second case of predictive validity was tested with a regression

analysis in which the ILT was used to predict scores on OLT1 and OLT2. The results of the regressions indicate that the scores on the in-level test can account for moderate amounts of variance in the predication of out-of-level test scores (see Table 12). In this study the predictive value of the in-level test (the hard test) was considerably greater than it was in the studies discussed earlier (cf. Boldt, 1968; Cliff, 1958; Levine & Lord, 1959). The reason might have been because the in-level or hard test was relatively easier for the subjects in this study than it was for the subjects in the earlier studies.

- - - - - - - - - - - - - - -

Insert Table 12 about here

- - - - - - - - - - - - - - -

Summary

The results of this study suggest that the support for the use of out-of-level testing with integrated special education students may need to be qualified. Analyses related to (a) the adequacy of vertical scaling, (b) the reduction of chance-level scoring, and (c) the reduction of guessed items suggest that the differences in scores may be accounted for by the students' levels of achievement and the test levels given. That conclusion is consistent with previous studies. In general, it does not appear that out-of-level testing of mildly handicapped special education students results in great reductions in the amount of measurement error. Nor does it appear that out-of-level testing results in substantial increases of internal consistency reliabilities. The assessment of the validity of out-of-level testing was limited. It appears that (a) the item validities do not improve significantly with out-of-level testing, (b) correlations of teacher ratings of student

achievement with obtained test scores are roughly equivalent across test levels, and (c) the substantial portions of the variance on the out-of-level tests which were accounted for by the in-level tests suggest that the in-level tests have fair predictive value. The results of this study suggest that out-of-level testing does not offer clear advantages over in-level testing in the routine assessment of special education students who have been integrated into the regular education program. It seems that perhaps general concerns about the reliability of in-level tests should not guide the decision to test out-of-level. Rather, careful attention to the content of each test level and its congruence with the instructional program should guide the decision to test in- or out-of-level. Other considerations would be the degree to which out-of-level testing would complicate the evaluation procedure, and the degree to which those persons responsible for the implementation of the evaluation were committed to it. Considering the logistical problems and extra expenses which are likely to accompany efforts to implement out-of-level testing evaluations, it seems that one should hesitate before deciding to use the procedure. The results of this study suggest that although out-of-level testing would not be inappropriate in the evaluations of special education programs for students with mild learning handicaps, it probably would not yield better quality data.

Limitations and Recommendations

There were several limitations to the generalizability of the results of this study. Most of them were related to the sample used in the study. Other limitations were related to the test and the methodology.

Sample Size

The sample used in this study was quite small (N = 58). It was, however, comparable in size to that of other studies (Crowder, 1978; Powers, 1978; Slaughter & Galles, 1978). A larger sample would have had the advantage of reducing the amount of error that had to be dealt with even further. A large sample also would have allowed for the analyses of other variables such as: achievement level, level of integration into the regular program, categorical labels, sex, and grade levels.

Achievement level. Very few of the students in this study scored below the chance-level on the in-level test--21.4 percent in the fifth grade and 13.3 percent in the sixth grade. Since they were all identified as being below the 20th percentile or at least two years below grade level in reading, the students would have been candidates for out-of-level testing under Title I guidelines. Despite their eligibility for such evaluation, the students evaluated in this study apparently were not extremely low achievers. That is, a substantial proportion of them did not score lower than would have been expected for a group of Title I students. The average percentile rank for the fifth graders was 22, and for sixth graders it was 23. The achievement levels may have been influenced by two factors related to the conditions of the study. First, the students were tested during the spring on test levels which were appropriate for fall and spring evaluation. Testing during the spring with those test levels meant that the students were evaluated on a relatively easier test than if they had been tested during the fall. Second, it is likely that the students in this study and the students in other studies differed considerably in their proficiency in English language skills. Although none of the studies gave detailed descriptions

of their samples, it is quite probable that there were substantial proportions of bilingual students in studies conducted in large metropolitan areas (e.g., Ayer & McNemara, 1973; Mendro, 1977; Wick, 1977; Wick & Ward, 1977) or in the southwestern United States (e.g., Mendro, 1977; Slaughter & Galles, 1978). It is recommended that future studies of the test-taking behavior of low achieving students should use test levels during their earliest appropriate testing period. In the case of this study a fall testing with the same test levels would have been preferred. It is further recommended that bilingual students be excluded, analyzed separately, or carefully accounted for in future research.

Level of integration. Data were gathered upon the level of integration of the special education students into the regular classroom programs. There was a great deal of variability in the amount and type of integration. From comments by teachers and administrators, it was decided that the amount and type of integration were more apt to be related to placement options and local school policies than to the students' academic statuses. It is recommended that schools with a well-developed range of services such as itinerant, resource, and part-time programs should be included in any further research on out-of-level testing.

Categorical labels. Most of the students in this study were labeled as LD. Since school psychologists typically categorized students with IQ's below 85 as EMR and usually recommended self-contained placements, there were very few EMR students in this study. Although it may be hard to find schools with better developed identification and placement criteria, they should be sought as sites for future research.

Sex. Only males were included in this study. If sufficiently large samples of females can be obtained, they should be included in future studies.

Grade levels. In this study only fifth and sixth graders were sampled. There were several reasons for imposition of that restriction. The first was that the levels appropriate for these grades had very similar number of items. The second reason was that for some analyses the grades could be legitimately combined. The third reason was that standardized achievement tests for elementary students are more apt to be sensitive to rather small change in difficulty than the tests considered appropriate for secondary students. A thorough investigation would require sampling mildly handicapped students at several different grade levels.

Representativeness

It did not appear that the sample of integrated special education students in this study was particularly unusual. However, self-selection must be considered a possible threat to the internal validity of the study because parental permissions were required. In some schools less than 30 percent of the parents returned permission slips to allow their children to participate in this study.

Tests

The CAT reading subtest was the only test used in this study. Different results might have been obtained if a test from another publisher or if another curriculum area subtest had been used. Arter (1980) reported that the different methodologies which have been used in out-of-level testing studies may have obscured any influences which might have been attributable to particular tests. Loyd (1980) reported that the

estimates of error for the math subtests decreased with appropriate out-of-level testing. The estimates of error for a less curriculum dependent subtest, Language, were not as sensitive to changes in test levels.

Design

The design of the study was another possible limitation to the generalizability of the results. Because of the small sample size, it was necessary to use the repeated measures design. It did not appear that maturation--the most probable threat to internal validity--was present. In this case the major limitation of the repeated measures design was that the analyses of differences between correlations presented major problems. A large sample, though very difficult to obtain, would have allowed the use of an independent group design.

REFERENCES

Arter, J. Functional level testing. Unpublished manuscript, 1980. Northwest Regional Educational Laboratory, Portland, Oregon.

Ayer, J. E., & McNamara, T. C. Survey testing on an out-of-level basis. Journal of Educational Measurement, 1973, 10, 79-84.

Barnes, J. A report on individualized testing in Atlanta. In Putting current ideas in measurement to work in the schools. Proceedings from the Thirteenth Annual Houghton Mifflin Measurement Conference, Iowa City, Iowa, 1977.

Birman, B. F. Problems of overlap between Title I and P. L. 94-142: Implications for the federal role in education. Educational Evaluation and Policy Analysis, 1981, 3, 5-19.

Boldt, R. F. Study of linearity and homoscedasticity of test scores in the chance range. Educational and Psychological Measurement, 1968, 28, 47-60.

Buros, O. K. The eighth mental measurements yearbook (Vol. I). Highland Park, N.J.: Gryphon Press, 1978.

Cliff, R. The predictive value of chance-level scores. Educational and Psychological Measurement, 1958, 28, 607-613.

Crowder, C. R. Relation of out-of-level testing to ceiling and floor effects on third and fifth grade students. Paper presented at the meeting of the American Educational Research Association, Toronto, 1978. (ERIC Document Reproduction Service No. ED 155-212)

Hallahan, D. P., & Kauffman, J. M. Introduction to learning disabilities: A psycho-behavioral approach. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1976.

Helmstadter, G. C. Principles of psychological measurement. New York: Appleton-Century-Crofts, 1964.

Henrysson, S. Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Kelly, T. Interpretation of educational measurements. Yonkers, N.Y.: World Book, 1927.

Levine, R., & Lord, F. M. An index of the discriminating power of a test at different parts of the score range. Educational and Psychological Measurement, 1959, 19, 497-503.

Loyd, B. H. Functional-level testing and reliability: An empirical study. (Doctoral dissertation, University of Iowa, 1980.) Dissertation Abstracts International, 1981, 42, 182A. (University Microfilms No. 8114281)

Mendro, R. A report on individualized testing in Dallas. In Putting current ideas in measurement to work in the schools. Proceedings from the Thirteenth Annual Houghton Mifflin Measurement Conference, Iowa City, Iowa, 1977.

Meyers, C. E., MacMillan, D. L., & Yoshida, R. K. Correlates of success in the transition of MR to regular class. (Final Report OEE 0-73-5263.) Washington, D.C.: Bureau of Education for the Handicapped, 1975. (ERIC Document Reproduction Service No. ED 116 441)

Powers, S. Implications of out-of-level testing for ESEA Title I students. Paper presented at the meeting of American Educational Research Association, Toronto, 1978. (ERIC Document Reproduction Service No. ED 164 441)

Pelavin, S. H., & Barker, P. A study of the generalizability of the results of a standardized achievement test. Paper presented at the meeting of the American Educational Research Association, San Francisco, 1976.

Slaughter, H. B., & Galles, E. J. Will out-of-level norm referenced testing improve the selection of program participants and the diagnosis of reading comprehension in ESEA Title I programs? Paper presented at the meeting of the American Educational Research Association, Toronto, 1978. (ERIC Document Reproduction Service No. ED 155 213)

Stewart, B. L. Interlevel articulation and Title I evaluation. Paper presented at the meeting of the American Educational Research Association, Boston, 1980a.

Stewart, B. L. Discussion: Assessing the reliability and validity of chance-level scores. In G. Echternacht (Ed.), Measurement aspects of Title I evaluations: New directions in testing and measurement. San Francisco: Jossey-Bass, Inc., 1980b.

Technical bulletin 2: California achievement tests. Monterey, CA: CTB/McGraw-Hill, Inc., 1978.

Wick, J. W. A report on individualized testing in Chicago. In Putting current ideas in measurement to work in the schools. Proceedings from the Thirteenth Annual Houghton Mifflin Measurement Conference, Iowa City, Iowa, 1977.

Wick, J. W., & Ward, F. Testing students at functioning reading level: A two-year report from Chicago. Unpublished paper, February 14, 1977. Board of Education, City of Chicago, Department of Research and Evaluation, 2021 N. Burling Street, Chicago, Illinois 60614.

Yoshida, R. K. Out-of-level testing of special education students with a standardized achievement battery. Journal of Educational Measurement, 1976, 215-221.

jgm4/1

Table 1
Summary of Demographic Data

Grade	N	Race (percent) Black	White	Categorical Labels (percent) ED	LD	EMR	Level of Integration (a) (percent) 1	2	3
5	28	21.4	78.6	3.6	96.4	0.0	57.1	39.3	3.6
6	30	30.0	70.0	10.0	86.7	3.3	60.0	16.7	23.3
Total	58	25.9	74.1	6.9	91.3	1.8	58.6	27.5	13.8

Grade	N	Age $\bar{X}$	Age SD	IQ $\bar{X}$	IQ SD	Estimated Reading Level (b) (GE Scores) $\bar{X}$	Estimated Reading Level (b) (GE Scores) SD	Hrs. per day in special ed. program $\bar{X}$	Hrs. per day in special ed. program SD
5	28	11.8	0.89	89	10.6	3.5	0.7	1.6	1.1
6	30	12.8	0.79	87	14.2	3.7	0.9	1.8	1.2
Total	58	12.3	0.96	88	3.6	3.6	0.8	1.7	1.2

[a]Levels of Integration into Regular Education Programs

1 = resource room--receiving instruction in reading in both the regular and special education class.

2 = part-time special class (a)--receiving reading instruction in special education class only, but does participate in some content area subjects with the regular class (e.g., social studies, math, science, or health).

3 = part-time special class (b)--receiving reading instruction in special education class only. Regular class participation is limited to non-academic subjects (e.g., art or physical education).

[b]Teacher estimates of reading level.

Table 2

Results of Two-way ANOVA with Repeated Measures: Mean Vertical Scale Scores Between Grades and Test Levels

Source	s.s.	df	m.s.	F	Significance
Between Grades	28690.063	1	28690.063	4.09	$p < .05$
Between Test Levels	21302.142	2	10651.071	3.33	$p < .05$
Interaction	964.809	2	482.404	.15	NS
Residual	358064.363	112	3197.003		

Table 3

Vertical Scale Score Means Standard Deviations and Standard Errors of Measurement by Test Levels and Grades

	Grade					
Test Level	5			6		
	$\bar{x}$	SD	SE_M	$\bar{x}$	SD	SE_M
ILT	403.3	115.3	$\pm$106.1	431.8	45.1	$\pm$34.2
OLT[1]	395.2	32.2	$\pm$26.8	424.7	46.7	$\pm$29.7
OLT[2]	381.7	27.7	$\pm$19.3	400.7	43.6	$\pm$23.9

Table 4

Kuder-Richardson-20 Reliabilities
by Grades and Test Levels

Grade	n	Test Level		
		ILT	OLT[1]	OLT[2]
5	28	.788	.827	.879
6	30	.859	.899	.925

Table 5

Frequencies Table for Cochran Q Test
Below 25 Percent Correct

Percent Correct	Test Level		
	ILT	OLT[1]	OLT[2]
	5th Grade		
≤ 25%	6	5	0
≥ 26%	22	23	28
Total	28	28	28
	6th Grade		
≤ 25%	4	3	2
≥ 26%	26	27	27
Total	30	30	29

Table 6

Results of Cochran Q Test for Significance of Changes in the Proportions of Raw Scores below 25 Percent Correct

Grade	n	Q-value	df	Significance
5	28	6.89	2	< .05
6	29	.86	2	N.S.

Table 7

Results of McNemar Follow-up for Paired Comparisons of Proportions of Raw Scores below 25 Percent Correct

			Paired Comparisons		
Grade	n	df	ILT-OLT[1]	ILT-OLT[2]	OLT[1]-OLT[2]
5	28	2	NS	$p < .05$	$p < .05$

Table 8

Frequencies and Proportions of Items with Non-significant Chi-square Values for Alternative Responses by Grades and Test Levels

		Test		
Grade		ILT	OLT[1]	OLT[2]
5	frequency	36	33	05
	proportion	.51	.47	.07
6	frequency	35	26	20
	proportion	.50	.37	.29

Table 9

Follow-up Analyses of Paired Comparisons of the Proportions of Guessed Items by Grades and Test Levels.

Contrast	Confidential Interval	Lower Limit	Upper Limit	Significance
		Grade 5		
ILT-OLT[1]	.043 ± .207	-.164	.250	NS
ILT-OLT[2]	.443 ± .165	.278	.608	$p < .05$
OLT[1]-OLT[2]	.400 ± .164	.236	.564	$p < .05$
		Grade 6		
ILT-OLT[1]	.129 ± .204	-.075	.332	NS
ILT-OLT[2]	.214 ± .197	.017	.412	$p < .05$
OLT[1]-OLT[2]	.086 ± .194	-.108	.279	NS

Table 10

Medians of Point-biserial Distributions
by Grades and Test Levels

		Test Level		
Grade	ILT[a]	ILT	OLT[1]	OLT[2]
5	.51-.56	.25	.25	.31
6	.50-.53	.31	.34	.40

[a]Medians for in-level test for the standardization sample (Technical bulletin 1, 1978).

Table 11

Spearman-rank Coefficients for Correlations between Teacher Estimated Grade Equivalents and Vertical Scale Scores across Test Levels (Combined Grades)

n	ILT	OLT[1]	OLT[2]
58	.5646	.6751	.5310[a]

Note. [a] based upon n = 57.

Table 12

Results of Simple Regression Analyses Where In-Level Test-performance Predicted Performance on Out-of-level Tests

Prediction	r	Adjusted r^2	F	df	Significance
OLT[1] from ILT	.489	.23	17.256	1,55	$p < .05$
OLT[2] from ILT	.403	.15	10.679	1,55	$p < .05$

Table 13

Frequencies for McNemar Follow-up Analyses of the Cochran Q Test of Proportion below 25 percent Correct. (5th Grade Only)

	ILT		
a. OLT[1]	≤ 25%	≥ 26%	Total
≤ 25%	2	4	6
≥ 26%	3	19	22
Total	5	23	28

	ILT		
b. OLT[2]	≤ 25%	≥ 26%	Total
≤ 25%	0	6	6
≥ 26%	0	22	22
Total	0	28	28

	OLT[1]		
c. OLT[2]	≤ 25%	≥ 26%	Total
≤ 25%	0	5	5
≥ 26%	0	23	23
Total	0	28	28

ss5/1