DOCUMENT RESUME

ED 169 126 TM 008 648

AUTHOR	Roberts, A. Oscar H.
TITLE	Out-of-Level Testing. ESEA Title I Evaluation and Reporting System. Technical Paper No. 6.
INSTITUTION	RMC Research Corp., Mountain View, Calif.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	Oct 76
NOTE	11p.
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Achievement Tests; *Complexity Level; Elementary Secondary Education; *Grade Equivalent Scores; Norm Referenced Tests; Scores; *Scoring; Standardized Tests; Student Testing; Testing Problems; *Test Interpretation; Test Reliability; *Test Selection
IDENTIFIERS	*Out of Level Testing

ABSTRACT

For Title I evaluations, it may be appropriate to test out-of-level; that is, to override publisher's recommendations concerning the difficulty, length, and content appropriate for a particular grade. It is seldom necessary, however, to move more than one grade down. If the mean is substantially higher than the median, then some pupils will have encountered the floor of a test and an easier level of tests should have been chosen. The ceiling of most tests becomes a handicap when three-quarters of a group can answer the most difficult items correctly. In this case, the mean is substantially lower than the median and a more difficult test should have been chosen. In general, the level of a test is suitable when the raw score of the group is equal to or above a third of the maximum score, and somewhat less than three-quarters of the maximum. In norm-referenced evaluations out-of-level testing is possible with most standardized achievement tests because they provide tables for relating raw scores on out-of-level tests to in-level percentile norms. (CP)

OUT-OF-LEVEL TESTING

ESEA Title I
Evaluation and Reporting System

Technical Paper No. 6

A. Oscar H. Roberts

October 1976

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

RMC Research Corporation
Mountain View, California

When test publishers collect material for tests, they decide the range of ability they want to cover, the number of items to have at each level of difficulty, the length of time the tests should take, and the suitability of the content for the age group at which the tests will be aimed. Inevitably some of these goals conflict with one another so that test design becomes a matter of compromises.

For example, a test of reading ability could be designed that would take only an hour to give, and would be usable at all grade levels. That would mean that first-grade students would be profitably occupied for about five minutes and frustrated the rest of the time, while twelfth-grade students would be overcome with boredom. Or again, a first grader could take some pride in reading "The cat sat on the mat" while a fourth-grade student, even though a very poor reader, might feel insulted by the choice of content.

To overcome these problems, tests are usually constructed with different levels, each level presumably suitable in terms of both content and difficulty for children of specific ages or in specific grades. The wider the age/grade band covered by a particular level, the more likely it is that either the difficulty or the content (or both) will be poorly matched to pupils at the upper and lower ends of the distribution. On the other hand, focusing test levels on too narrow a band of difficulty or content produces other kinds of problems, including that of simply being confusing.

REASONS FOR OUT-OF-LEVEL TESTING

The publishers of the major achievement tests have highly qualified personnel, up-to-date techniques, and years of experience developing and scaling tests. While it would be hard to improve upon the compromises they have made with respect to test level difficulty, length, and content, special conditions may sometimes make it desirable to override the publisher's recommendations as to which level of a test should be used at a particular grade. This circumstance is likely to arise in Title I settings when students with the poorest performances are tested. It should be noted that a lower level test will not be needed for every Title I group. A look at some of the factors that are considered by test designers might be useful when trying to decide whether to test out of level.

No test measures exhaustively; it samples skills or abilities and it does so for the best and the poorest students simultaneously. Data from samples, as in the case of opinion polls, can yield quite accurate predictions, but the approximations are poorer when samples are smaller. Thus we need to ensure that the proportion of test material on which students can profitably spend time and effort does not drop too low--as it would if the test were either much too difficult or far too easy.

Floor Effects

What is the optimal proportion? To answer this question first consider multiple-choice instruments; they have enough advantages to make them the best choice for standardized tests, but they do have some unavoidable disadvantages. One of these is that even when a group of students is quite out of its depth, erroneous thought processes or guessing can yield apparently "interpretable" scores. For example, if 100 students

completed a 32-item, four-choice test by guessing alone, the average score would be about eight, and the scores would range from a low of about three to a high of around thirteen. This range that occurs as a result of guessing is a serious problem because it severely reduces the reliability of a test.

From the example above, it can be seen that it is possible to encounter the "floor" of a test even though no students have scored zero. Long before the average raw score of a group is at chance level (the score equal to the total number of questions divided by the number of alternative answers to each question), some pupils in the group will have encountered the floor and an easier level of test should have been chosen.

If you suspect that you may have encountered a floor effect, a convenient check is to compare the mean with the median of the scores. If the mean is substantially higher than the median (by about a third of a standard deviation) then your suspicions are very likely confirmed.

Ceiling Effects

Unfortunately, there is also the other extreme--too easy a test. If we seek to avoid possible "floor" effects by choosing a lower level of test, we could bump up against the "ceiling." Once again this can occur even if no one achieves the highest possible raw score since carelessness and accident are more likely to occur if the test items are too easy. It is more difficult to find a way of setting limits here, but, in practice, the ceiling of most tests becomes a serious handicap when three-quarters of a group can answer the most difficult items correctly.

If you suspect that you may have encountered a ceiling effect, a convenient check is to compare

the mean with the median of the scores. If the mean is substantially lower than the median (by about a third of a standard deviation), then your suspicions are very likely confirmed.

DETERMINING THE APPROPRIATE LEVEL

In most instances the level of a test is suitable when the mean raw score of the group is equal to or above a third of the maximum score, and somewhat less than three-quarters of the maximum. The highest reliability of a test is achieved when the students, on the average, get slightly more than half the items correct. However, unless previous test scores are available as guidance, one has to depend upon teaching experience and judgment to select the correct test levels. It should seldom, if ever, be necessary to move more than one level down; and even that is likely to be unnecessary when, for example, the group comes from grade 4 and the test is suitable for grades 3 and 4.

Test publishers try to avoid the occurrence of ceiling and floor effects and to construct their tests so that the median score at the appropriate grade level is well above half the number of items in the test. Thus, for an average class, students are more likely to score close to the ceiling of the test than close to its floor. If the same test is used at a higher grade level, the trend is increased.

It can be seen that, if too low a test level is used, scores will be artificially depressed. If this occurs on the posttest and not on the pretest, gains will also be depressed. If the test ceiling is encountered only on the pretest and not on the posttest (because, presumably, the level was changed), gains will be spuriously inflated.

It is never proper to do out-of-level testing simply to give pupils the experience of success--especially when the practice could result in encountering test ceilings. On the other hand if, as in the Stanford Achievement Tests, many of the levels are intended for use at the end of one grade and at the beginning of the next only, it would be quite reasonable to use the lower level for both pre- and posttest in a fall-spring design, even where the fall testing was in-level and the spring necessarily out-of-level.

INTERPRETING SCORES FROM OUT-OF-LEVEL TESTING

In norm-referenced evaluations, out-of-level testing is possible only with tests that provide an expanded standard score scale. This scale allows the raw scores on the out-of-level test to be related to the in-level percentile norms. Most of the major standardized achievement tests presently have this type of scale, but the conversions which must be made will depend upon whether the publisher has provided raw-score-to-percentile, or standard-score-to-percentile conversion tables. In either case, the goal is to determine the percentile rank (or NCE) that would, in theory, have been obtained if the appropriate level of test had been used.

Tests with Expanded-Standard-Score-to-Percentile Conversion Tables

Some achievement tests convert raw scores to expanded standard scores for each test level, and then provide a separate table converting the expanded standard scores to percentiles for each grade and time of year. Tests requiring these conversions include the Iowa Test of Basic Skills, the Metropolitan Achievement Test, the Sequential

Tests of Educational Progress II, and the SRA Achievement Series. (Note that the expanded standard score scales may have different names in different tests, e.g., "standard scores," "scale scores," or "converted scores." The name does not always indicate whether the scores are expanded to cover different grade levels. For this information, refer to the RMC Technical Paper No. 5, entitled Characteristics of Eight Commonly Used, Nationally Normed Tests, or to the test publishers' manuals.)

For tests that convert expanded standard scores to percentiles, use the following procedure: convert each student's raw score on the level of the test which was administered to its corresponding expanded standard score and compute the average. Then, convert the average expanded standard score to a percentile or NCE using the tables for the "appropriate" test level.

Suppose we have a group in grade 4 for which the Green Level reading test is nominally recommended. Instead we used the Blue Level which is one level lower. Assume that the testing was done in the fall at a time which corresponded to an empirical normative data point. To obtain the appropriate percentile (or NCE) value, we should:

1. Convert each student's raw score to an expanded standard score using the table for the Blue Level.

2. Find the average of these standard scores. (Assume it was 64.)

3. In the manual for the Green Level, use the standard-score-to-percentile conversion table for beginning of 4th grade and find the percentile (or NCE) that corresponds to the standard score. In this example a standard score of 64 might correspond to the percentile rank of 39.

Tests with Raw-Score-to-Percentile Conversion Tables

Instead of the conversion tables just discussed, some tests provide tables that convert raw scores to expanded standard scores, and raw scores to percentile ranks. These tests include the California Achievement Tests, the Comprehensive Tests of Basic Skills, and the Stanford Achievement Test. For such tests, we would convert the raw scores for the level of the test which was administered to expanded standard scores and find their mean. This value we would then take to the tables for the "appropriate" level, and find the corresponding, in-level raw score. Finally, in the appropriate table for this higher level, we would find the percentile rank (or NCE) for that raw score.

Suppose, for example, that we posttested a grade 7 group in spring using the Orange Level of a reading test when the Red Level was the recommended one. To find the appropriate percentile (or NCE) we should:

1. Convert the raw score of each student to the corresponding expanded standard score, using the Orange Level raw-score-to-expanded-standard-score conversion table.
2. Find the average of these standard scores. (Assume it was 423).
3. In the manual for the Red Level, use the raw-score-to-standard-score conversion table to determine the in-level raw score corresponding to the mean standard score. Our value of 423 might correspond to a raw score of 33.
4. Finally, in the same manual (Red Level), using the end-of-7th-grade tables, convert this raw score to a percentile (or NCE). In our example, the raw score of 33 might have a corresponding percentile value of 28.

It is easy to see that this process is similar to the one used for tests that convert expanded standard scores to percentiles, but one extra step is needed to go from the expanded standard score to the appropriate in-level raw score.