HAS ITEM RESPONSE THEORY INCREASED THE VALIDITY OF ACHIEVEMENT TEST SCORES?

CSE Technical Report 302

Robert L. Linn

University of Colorado

UCLA Center for Research on Evaluation, Standards, and Student Testing

May, 1989

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

This report was written as part of a research project sponsored by the UCLA Center for Research on Evaluation, Standards, and Student Testing (CRESST). It was prepared for presentation at the Annual Meeting of the American Educational Research Association, San Francisco, March, 1989. This report will appear as a chapter in *Dimensions of Thinking and Cognitive Instruction* (B.F. Jones and L. Idol, editors), to be published by Lawrence Erlbaum Associates.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

IRT and the Question of Validity

Item response theory is probably the most important technical development in the field of measurement in recent years. On the other hand, the most important question regarding any measure concerns the validity of the uses and interpretations of the scores. Relatively little attention, however, has been given to the influence that increased use of IRT has had on the validity of test scores. The purpose of this paper is to address the question posed in the title. Has IRT increased the validity of achievement test scores?

As will be seen, the proposed answer to the question will not be a simple "yes" or "no". Rather, it is argued that IRT has made some important contributions. It has opened up new possibilities that were beyond our reach without the power of IRT. It has also raised a host of new questions and forced us to think about a number of measurement issues in new and more penetrating ways. Often, the new questions and issues that are posed lead to the need for refinements and extensions of the theory and there have been some important new developments in these regards in the past few years.

On the other hand, the IRT models for responses to test items, like any other model are only approximations. Wainer and Thissen (1987) made this point in the introduction to their article, "Estimating Ability with the Wrong Model" with the following quotation that they attributed to John Tukey. "All theories are wrong. Its just that some are easier to disprove than others" (quoted by Wainer & Thissen, 1987, p. 339). Thus, it should not be surprising, that given enough data and sufficiently powerful analytical techniques, that we can find ways in which any IRT model is wrong. More important, however, are questions such as the following. What effects do the differences between the model and reality have on the validity of interpretations and uses of a measure? What new insights do we gain about reality by virtue of this comparison? Are there better alternatives available for particular purposes? If not, what needs to be done to develop such alternatives?

In addition to questions about the adequacy of the model for particular applications with a given set of data, it is important to also address the question by starting with a consideration of the purposes of educational measurement, or more specifically, achievement testing. The validity of a measure needs to be evaluated in light of the purposes of measurement and the ways in which measures are used and interpreted. It is within this context that we can then look at the degree to which IRT contributes to validity and enhances or impedes the accomplishment of those purposes.

A variety of purposes for achievement tests can be identified. Commonly mentioned are such seemingly diverse purposes as accountability, certification, and the identification of student strengths and weaknesses. Despite the apparent diversity, however, the specific purposes all relate in one way or another to the global goal of improving learning. Although no attempt will be made to provide definitive conclusions about the degree to which IRT has increased the effectiveness of achievement tests in terms of this global goal, it is useful to keep the goal in mind.

Before moving to some specific educational measurement issues where the question of the contribution of IRT can be addressed more concretely, it is useful to review briefly notions of validity that will be used to evaluate the contributions in particular applications. Some of the primary expected advantages of IRT also need to be highlighted and linked to applications with achievement tests. After this general introduction four specific applications will be discussed. The specific applications to be considered are (1) the construction of scales for achievement tests, (2) the selection of items for achievement test forms, i.e, test construction, (3) the construction of customized tests, and (4) the investigation of the influence of instruction on achievement tests.

Validity

Messick (1988) conceives of validity as a broad, but unitary concept that encompasses "two interconnected facets ... "One facet is the source of justification of the testing, being based on appraisal of either evidence or consequence. The other facet is the function or outcome of testing, being either interpretation or use" (p. 20). The function or outcome facet is the more familiar of the two. Recommendations that evidence be accumulated concerning both test interpretation and test use are a standard part of recent discussions of validity (e.g., American Psychological Association, 1985).

Although Messick's concern for appraising the consequences of testing is consistent with the views of other major theorists (e.g., Cronbach, 1980; 1989), his source-of-justification facet is a less familiar way of classifying validity than the distinction between use and interpretation. The need to make a judgment about the consequential basis as well as the evidential basis of test use and interpretation greatly expands the range of validity considerations.

A broad view of validity such as the one articulated by Messick, suggests that questions about the degree to which IRT has increased validity of achievement tests need to take a variety of forms. Certainly, we need to be concerned with evidence supporting inferences that are made from the scores and the uses of the scores. In addition, however, we need to ask about consequences. For example, are the intended purposes of achievement tests accomplished better because of IRT? What effect, if any, has IRT had on unintended consequences?

Some Advantages of IRT

To start, it is useful to note some of the expected advantages of IRT. Its advantages derive largely from the promise of invariance, or as was more colorfully described by Wright (1969), person-free item calibration and item-free person measurement. The promise of item-free person measurement is central to the possibilities of item banking, computerized-adaptive testing, and the development of customized tests. Person-free item parameter estimates provide the basis for certain approaches to test equating and it is this property that is challenged in studies of item bias. Questions about item parameter invariance are also central to studies of the differential effects of instruction on performance on achievement test items.

Of course the notion of item-free ability estimates is not meant to suggest that equally good estimates of an individual's ability can be obtained with any set of previously calibrated items. The precision of the estimates will depend on the number of items and their parameters. However, IRT provides the basis for estimating the amount of information, or conversely, the magnitude of the standard error of ability estimates provided by any given set of items at each point along the ability scale. Thus, it provides the basis for designing tests that have desirable properties for specific purposes (e.g., reliable measurement over a wide range of ability or measurement with maximum precision at a given point on the scale). With this background, I'll now turn to a consideration of four specific applications of IRT, starting with the construction of scales.

Scales

One of the uses of IRT with achievement tests is, of course, the construction of scales that span the levels covered by the test battery. When the assumptions of the IRT model hold, IRT scaling has a number of advantages. Unlike Thurstone scaling which has been used frequently with achievement test batteries, IRT scaling does not require the assumption that achievement is normally distributed within-grade. Another major advantage, as Burket (1984) has pointed out, "stems from the fact that IRT models the probability of the correct response of an examinee to an item, and therefore permits the interpretation of a test score in terms of what that score implies about the examinee's ability to perform. Given an appropriately

chosen set of calibrated benchmark items, this means that true criterion referencing of test scores in possible" (p. 15).

The criterion referencing that Burket was referring to has been used extensively with the IRT-based NAEP scales that have been used since the 1984 reading assessment. The selected benchmark items show the kind of problems that students at a given scale score, say 250, have a high probability of answering correctly whereas students with a lower scale score, say 200, are unlikely to answer correctly. Such benchmark items play an important role in interpreting the scale.

Although there is widespread agreement that IRT offers advantages for scaling when the model holds, there has been considerable debate about the use of IRT to scale achievement tests. Hoover (1984), for example, acknowledged that IRT has advantages, "but only if some highly restrictive conditions are met" (p. 17). He then went on to argue that the conditions are generally not met for standardized achievement tests, noting in particular, that "the assumption of unidimensionality is almost sure not to hold across the levels of an achievement battery in most test areas" (p. 17).

Although concerns about unidimensionality are enough to produce debates about the use of IRT to place the levels of an achievement test on a common scale, particular properties of the scale that resulted when IRT scaling was first used with the CTBS probably did more than anything else to call attention to the issue. Several reviews of the CTBS/U (e.g., Hoover, 1984; Linn, 1985; Shepard, 1985) noted that the scale indicated that the variability of student achievement tended to decrease as grade level increased. This decreasing variability result stood in sharp contrast to the more familiar Thurstone scales which tend to show some increase in variability with grade level. The pattern also seemed to conflict with some findings from differential psychology (e.g., Anastasi, 1958) and concepts such as Campbell's (see, for example, Cook & Campbell, 1979) fan-spread hypothesis, a kind of rich get richer theme, which posits that performance becomes more variable across the grades.

Hoover (1984) addressed the issue of decreasing scale variability in some detail (see also Burket, 1984, for a response). Yen (1985, 1986) also addressed the issue and provided a possible explanation for the tendency for the scales to shrink. Scale properties continue to be an issue, however, as is apparent in a recent exchange of papers by Phillips and Clarizio (1988a, 1988b), Yen (1988), and Hoover (1988).

Although there is still substantial disagreement about the appropriateness of certain IRT scales, there is general agreement on at least two points. So it is useful to set those aside at the beginning. First, while it has sometimes been claimed that equal interval scales are produced by both Thurstone scaling and IRT scaling, it is obvious that they can't both be equal interval when one scale suggests that above-average students tend to develop at a faster rate than below-average students, whereas the other scale suggests just the opposite. It is wiser to avoid a claim of a true equal interval claim in both cases, however, than to accept the claim in one case, but not the other. As Burket (1984) has noted, in both cases such claims "fall apart under critical scrutiny" (p. 15).

A second point about which there seems to be general agreement is that simply because an IRT scale has different properties than an those of the more familiar Thurstone scales, it does not necessarily follow that there is something wrong with the IRT scales. The fault, if there is one, may lie with the Thurstone scales. Alternatively, they may both be useful for particular purposes. To again quote Burket (1984), it is important to recognize that each scale "is based on a model whose predictions depend on the metric" (p. 15).

The concern about a shrinking scale comes, in part, because it runs counter to a widely held belief. According to Hoover (1984), for example, "One of the most widely held beliefs regarding the educational development of students is that above-average students develop at a faster rate than below-average students" (p. 11). Of course, a widely held belief can also be an erroneous belief. More importantly, however, the belief may be correct for some developed

abilities but not others and this raises fundamental validity questions. What is the construct that is being measured by a given test of an achievement battery? Is it a unidimensional construct? How do students develop the ability that is being measured? What are the consequences of choosing one scale rather than another for the use and interpretation of the scores?

Unfortunately, we seem to have little solid basis for answering such questions. Yen (1985, 1986) has suggested that increasing multidimensionality with test level may cause the scale to shrink. More specifically, the idea is that the higher test levels include more complex items that require several different abilities to solve.

Figure 1 (see Appendix A) displays the IRT-based scale scores corresponding to selected percentile points (i.e., the 10th, 25th, 50th, 75th, and 90th) for four reading comprehension tests across several grade levels. The plot in the upper-left-hand corner for the CTBS/U Reading Comprehension test illustrates the shrinking scale property that led to the initial debate. The scale score needed to be at the 90th percentile changes very little from the spring of the 5th grade to the spring of the 8th grade. On the other hand, fairly substantial growth would be required over that same three year period for a student to maintain a standing at the 10th percentile.

As can be seen in the upper-right-hand corner of Figure 1, the tendency for the scale to shrink is even more marked for the CAT than for the CTBS. The remaining two reading scales shown in Figure 1 give a rather different perspective. The scale variability of the Stanford is fairly constant across time. The NAEP reading scale shows a small decrease in variability from grade 3 to grade 7, followed by a slight increase in variability from grade 7 to 11. There is no good basis for saying that one of the four plots corresponds better to how student reading ability develops than the others. The differences between the plots do suggest that the tests have rather different measurement properties and possibly that they are getting at rather different underlying constructs of reading achievement and are apt to lead to rather different inferences about student learning.

Figure 2 presents results for four mathematics tests in parallel fashion to the Figure 1 results for reading. In addition to the points made in reference to Figure 1 that might be repeated in reference to Figure 2, the contrast between the CTBS and the CAT is particularly worthy of note. The CAT displays considerably less scale shrinkage in math than is shown by the CTBS, whereas the converse is true in reading.

There are potentially severe limitations in using percentile points of observed score distributions which in this case are the scale scores based on IRT theta estimates. Yen (1983; 1986) has clearly described the pitfalls. As she has demonstrated, tau-equivalent tests that differ in difficulty can produce different observed score distributions "due to the tests having different amounts of error variance at different parts of the scale" (1986, p. 315). Yen's (1986) simulation also demonstrated that the average difference between true and estimated growth when a single test is used as both a pretest and a posttest fluctuates by percentile level as a consequence of the shift in the size of the standard error of measurement at different points along the scale. Depending on the choice of the test information function, the use of fixed percentiles might make it appear that low scoring examinees were growing slower, faster, or about the same as their counterparts with middle or high scores. Thus, it is important to consider the relative magnitude of the standard errors of measurement at different scale score levels.

Before considering the possible influence of different matches between standard error curves and the distributions of examinee achievement level it is worth noting another feature of Figure 2. In addition to the obvious difference in the amount of scale shrinkage on the CTBS and CAT math tests, another contrasting feature of those two graphs can be seen with careful study. This is the tendency for the within-grade variability to shrink from fall to spring on the CTBS while it tends to increase on the CAT. The latter point can be seen more clearly in Figure

3 where the differences between the 90th and 10th percentiles are plotted for the those two tests.

The within-grade scale shrinkage has recently been addressed in a study by Camilli (1988). Camilli simulated tests that might be used in both fall and spring and demonstrated that with maximum likelihood estimation of abilities the shrinkage could be predicted by the degree of match between the item difficulties and the examinee abilities at the two points in time. In particular, if the item difficulties are well matched to the student abilities in the fall and there is substantial growth between fall and spring the difficulties will match the spring abilities less well. As a consequence, the variability of the estimated abilities will shrink even though the true abilities were equally variable at both points in time.

One way of considering the match of items to abilities is through the use of the information function or the magnitude of the standard errors of measurement at different scale points. Based on Camilli's simulation it seems reasonable to expect that the standard error of measurement curve would suggest that the CTBS tests are pitched to provide better measurement across the student score range in the fall than in the spring, while the converse would be true for the CAT.

Since the contrast between the apparent decrease in the within-grade variability for the CTBS and the increase in the within-grade variability on the CAT is greatest for Grades 2 and 3, I will focus on those two grades in considering the standard errors of the tests. Table 1 (see Appendix B) shows the relative size of the standard errors of the CAT and CTBS Math Concepts and Applications tests for fall and spring of Grades 2 and 3 for scale scores corresponding to selected percentile ranks between 5 and 95.

As would be expected, both tests show smaller standard errors of measurement for scale scores corresponding to the 90th or 95th percentile in the fall than for the scale scores with equally high percentile ranks in the spring. Also as expected, the converse is true for low percentile ranks, where the standard error of measurement is smaller in the spring. It is not obvious, however, that overall there is a relatively better match of the CTBS to the student distribution of performance in the fall than the spring. Nor is it obvious that the CAT is better matched to the spring than the fall. If anything, just the opposite would seem to be the case at Grade 2.

The most apparent difference between the ratios of standard errors in Table 1 is that the CAT maintains a more nearly constant standard error over a wider range of the observed scale score distribution at Grades 2 and 3 than does the CTBS. This is most notable at Grade 3 where the standard error of measurement is no more than one and a half times its minimum value throughout the range of scale scores ranging from the 5th percentile in the fall to the 95th percentile in the spring. In contrast, the ratios for the CTBS are much more U shaped.

Clearly, we need to learn more about the factors that affect IRT scale properties on achievement tests. The match of difficulty or the shape of information functions in comparison to examinee distributions is worthy of additional investigation. More exploration with alternate estimation procedures also seems desirable. In the case of Camilli's simulation, he found that the within-grade scale shrinkage could be avoided by the use of empirical Bayes estimates. Since empirical Bayes estimates have been used for NAEP, perhaps that helps explain why the NAEP scale does not shrink.

As Yen (1986) concluded, "IRT does not offer a simple answer to the question of what is the best method of scaling educational achievement tests" (p. 322). What it has done, however, is to bring some ignored assumptions about all approaches to scaling achievement tests into the open for more careful examination. The debate about scale properties has called into question the validity of widely accepted inferences about the growth in student achievement that are dependent on the particular properties of scales. Increased awareness of the scale dependent nature of certain interpretations is itself a contribution that IRT has made to validity of inferences about achievement that are made from test scores regardless of

whether one uses scores based on Thurstone scaling, IRT scaling, grade equivalent scores, or some other scale.

Test Construction

The second application of IRT that to be considered is its use in the selection of items for achievement tests. Lord (1977) described a four-step procedure that was first suggested by Birnbaum to use IRT in the selection of items for a test. The procedure starts with the selection of a target information curve for the test. Items with information curves that "will fill the hard-to-fill areas under the target information curve" (p. 120) are then selected. Part-test information curves are computed for already selected items. The process of selecting items and computing part-test information functions is continued until a satisfactory approximation to the target information function is achieved.

This procedure is quite reasonable if the assumptions of the IRT model are satisfied. It is quite efficient, for example, for selecting a set of items that will yield a test with minimal standard errors of measurement over a range of ability that is of particular interest for the test. As was previously indicated, however, an IRT model is only an approximation and potentially important assumptions such as the assumption of undimensionality will not be perfectly satisfied in practice. Thus a reasonable question is how well an IRT test construction procedure such as the one outlined by Lord works for the purpose of constructing achievement tests?

Notably absent in description of the item selection procedure is any mention of the content of the items to be selected. Yet, for achievement tests the content specifications are generally considered to be of primary importance. The primacy of content has been emphasized by a number of authors as have concerns about a possible distortion of the measurement that may result from an over-reliance on statistics.

Anderson (1972), for example, argued forcefully against the use of item difficulty and item discrimination in the selection of items for an achievement test. Although Anderson's comments were made in the context of classical item analysis statistics, his concerns apply equally to an IRT test construction approach. In his view, the use of discrimination indices to select items not only redefines the achievement domain that the test is intended to measure in unspecified ways but is apt to result in more of an aptitude measure than an achievement measure. According to Anderson, "manipulating tests to control difficulty level and discriminating power tortures validity ..." (1972, p. 82).

Several other authors have expressed concerns similar to those raised by Anderson. Oscar Buros (1948, 1977), for example, argued on several occasions for the need to attend more carefully to content and rely less on item statistics in constructing tests. Anderson's concern that standardized achievement tests are more measures of aptitude than achievement is also shared by others. Willingham (1980), for example, concluded that "standardized achievement tests are probably too often saturated with aptitude" (p. 78) and Jones (1988) has recently identified this confounding as a key educational measurement problem.

In practice, of course, items for standardized achievement tests are not selected solely on the basis of item statistics. The IRT item parameters are used along with other information. Yen (1983), for example, noted that Automatic Test Selection computer program that was used to help select items for the CTBS forms U and V included "requirements for a minimum number of items for each category objective" (p. 131). The program also maximized an index of overall item quality that included ratings of model fit and item bias as well as item discrimination. In addition, editors "refined the selected test, focusing on content considerations in particular" (p. 131). Thus, to some extent the specter of purely statistical construction of achievement tests may be something of a strawperson. The extreme form of the practice is useful, however, in highlighting the criticality of content considerations for achievement tests.

Traub and Wolfe (1981) noted that there is a disparity between the goals of achievement testing and the concept of a unidimensional trait. They cautioned that efforts to satisfy the undimensionality assumption "will almost certainly restrict the range of achievement tested" (p. 383). However, defenders of the use of IRT for achievement tests counter with at least three arguments. First, it is noted that classical test development procedures treat achievement areas as essentially unidimensional when only a single score is reported for a set of items and by the procedures used to equate test forms. Zwick (1987), for example, has pointed out that "when less sophisticated methods, such as the summation of item scores, are applied, it is implicitly assumed that the items are measures of a single attribute. IRT merely formalizes this assumption" (p. 293). Thus it is not obvious that IRT restricts the range of achievement tested to a greater extent than the reliance on classical item discrimination indices. On the other hand, as has already been noted, the criticisms such as those made by Anderson apply to classical procedures as well as IRT-based procedures.

A second rejoinder to critics of the use of IRT in the construction of achievement tests is that IRT models are falsifiable. As Lord (1980) has noted, "it is possible to make various tangible predictions from the model and then check with observed data to see if these predictions are approximately correct" (p. 15). Improved procedures for assessing dimensionality such as Bock, Gibbons, and Muraki's (1985) full-information factor analysis are especially useful in this regard. Considerable room for debate remains, however, on issues of just how dominant the first major dimension needs to be and whether dimensionality is something that changes as a function of instruction.

Snow and Lowman (1988) argue, for example, that "a test might be unidimensional for novices because all problems are relatively novel for them and, thus, require the same general problem-solving skills, whereas experts might show different patterns of skill development on different types of problems" (p. 267). Snow and Lowman's suggestion goes beyond Yen's (1985) hypothesis that more difficult items are more complex and therefore more likely to lead to multidimensionality. Snow and Lowman's suggestion is concerned with the dimensionality of a single set of items administered to groups that are at different stages of development, or, for that matter, the same group of students before and after a period of learning. This issue of possible changes in dimensionality as the result of learning will be considered in greater detail below, for it seems to be a particularly critical issue for achievement testing.

A third response to the undimensionality concern is to work with increasingly homogeneous content domains. Although undimensionality may not hold for a relatively broad domain such as mathematical concepts and applications, that domain can be further and further subdivided until the assumption is satisfied. Relatively narrowly defined skill domains of the type described by Bock, Mislevy, and Woodson (1982) and Pandey and Carlson (1983) do much to reduce the concerns about dimensionality. The goal, as Mislevy (1983) described it is to identify "item domains sufficiently homogeneous with respect to content that all the items in a given domain would be similarly affected by changes in curricular emphasis" (p. 273). The scaling is then carried out within those item domains. Summary scores can be obtained by various combinations of the "indivisible curricular element scores" if desired, but that would be accomplished after the scaling.

There is considerable diversity in the definitions of domains that are treated as unidimensional in applications of IRT. In reading, for example, NAEP used a single scale for ages 9, 13, and 17. On the other hand, the California Assessment Program (CAP) has defined narrow scales for each grade and subject area. In reading, for example, there are 17 skills at grade 3 (Mislevy, 1988).

Mislevy (1988) has suggested that the choice of level of aggregation should depend, in part, on the purposes. Many, narrowly-defined scales provide a means of detecting small shifts in curricular emphasis, but may be more than are needed for other purposes. Mislevy goes on to conclude that "one scale per subject area is probably too few, but twenty is probably too many" (p. 179). Exactly when content differences can be safely ignored and when they

require detailed attention is difficult to determine. It is clear, however, that content considerations need careful attention regardless of the level of aggregation.

Customized Tests

The issues of content representation and how nearly unidimensional an item pool needs to be have been made more salient in the last few years by the development of customized tests. Customized tests may be constructed by allowing a state or district to select items from a previously calibrated item bank rather than administering a complete off-the-shelf, norm-referenced test. In theory, it should be possible to estimate achievement levels on the same scale that would be provided by the off-the-shelf test no matter what subset of items was selected. If the model is correct and enough items with appropriate item parameters are selected to provide reliable measurement, then the norms presumably would be applicable.

But the model is just an approximation. In practice, it is clear that it is unwise to construct customized achievement tests without carefully controlling the content covered so that the customized and norm-referenced tests have what Yen, Green, and Burket (1987) referred to as content equivalence. The importance of content representation to the results of a customized test has been demonstrated in studies by Allen, Ansley, and Forsyth (1987) and by Way, Forsyth, and Ansley (1989). In these studies simulated customized tests were constructed from off-the shelf standardized tests by deleting items in some of the content areas covered by the shelf test. For example, Way, Forsyth, and Ansley started with the 40 items on the sixth grade level of the Language Usage and Expression test of the ITBS and constructed a 22-item content-customized test by selecting the items in the four item content categories concerned with usage and deleting the 18 items concerned with expression. Similar content related deletions of items were made by Allen, Ansley, and Forsyth for the Quantitative Thinking subtest of the ITED and by Way, Forsyth, and Ansley for three other ITBS tests (Vocabulary, Visual Materials, and Mathematics Concepts).

To evaluate the impact of the use of content to select items for a customized test schools that had high proportion right scores in the content categories covered by the customized test relative to their overall proportion right scores were selected. These schools were intended to simulate what might happen if schools selected content categories that corresponded most closely to objectives that they emphasized while eliminating content categories deemed less important and given less emphasis. Based on their analyses, Way, Forsyth, and Ansley concluded "that if schools in this study were to use the ability estimates based on the respective content-customized tests instead of the full-length, tests, the significant majority of examinees would receive higher normative scores than they would if the full test were used" (p. 34). They went on to say that "apparently it would not be valid for the school to refer ability estimates based on the content-customized test to norms that were established on the basis of the full test" (p. 34).

Yen, Green, and Burket (1987) provided compelling evidence that content considerations are essential if valid normative information is to be obtained from a customized test. They compared the IRT b values obtained for the national norm group with those obtained for a local educational agency for a grade 5 Mathematics Concepts and Applications test. The scatterplot of the LEA and norm group b values clearly indicated that the item parameters were <u>not</u> sample-free. More importantly the plot showed that the content of the items was related to the shifts in the b values. In general, measurement items were relatively more difficult for the LEA students than for the national sample, whereas the converse was true of numeration items. As Yen, Green, and Burket noted, this calibration group by item content interaction could result in invalid normative estimates on a customized test "if the content of the customized test did not proportionally represent the content of the normed test" (p. 9).

Yen, Green, and Burket's results are of considerable importance for anyone considering the use of customized tests to obtain normative estimates. As they concluded, "if an LEA wants to customize a test with respect to content, it is likely that local instruction could differentially

affect performance on the customized test and the normed test, invalidating normative information based on trait estimates from the customized test" (p. 13).

The magnitude of the effect of using a content-customized test to obtain normative estimates can be dramatic. This is illustrated by the results shown in Figure 4 for a state that switched from an off-the-shelf, norm-referenced test to a customized test. Figure 4 shows the percentage of students by stanine score on a Grade 5 mathematics test for each of 4 years. For Years 1, 2, and 3, the results are based on the administration of a norm-referenced test. In year 4 a customized test which did not proportionally represent the content of the normed test was administered. As can be seen there was gradual improvement in the test scores in Years 2 and 3. The apparent increase in performance in Year 4, however, is extraordinary. The percentage of students with stanine scores of 8 or 9, for example, sky-rocketed from 13% in Year 3 to 36% in Year 4. These results suggest that the conclusions of Way, Forsyth, and Ansley, and of Yen, Green, and Burket about the likely invalidity of normative information based on content-customized tests need to be carefully heeded.

A major question that needs more attention is the degree to which proportional representation of content and the selection of items with appropriate item parameters are adequate safeguards for insuring the validity of normative information based on the trait estimates from a customized test. Matching as closely as possible in terms of content is clearly desirable for customized tests just as it is for the construction of alternate forms of norm-referenced test. But it may not be sufficient. Unlike an alternate form of a test, a customized test may differ from a normed test in other ways that could influence results. Test length, the context provided by criterion-referenced test items that may be included in the customized test but not used to obtain norms, and item position effects are issues that seem worthy of more study in this regard.

Instruction and Achievement

The concerns for content representation and the effects of content on performance by a state or district on a customized test provide a natural lead into my final topic, the degree to which variation in curricula and in instruction differentially affect performance of items assumed to measure a common dimension according to the IRT model. Questions regarding the sensitivity of achievement tests to differences in curricula and instruction, the importance of alignment between tests and curricula, and about the importance of the degree of match between what is taught and what is tested are, of course, not limited to achievement tests that rely on IRT. They are just as relevant for tests relying on different technologies. As is true of a number of other general measurement issues, however, IRT makes some of the issues more apparent because the assumptions of the models are more explicit. IRT also provides a framework for addressing the issues in more illuminating ways.

According to Lord (1980), "The invariance of item parameters across groups is one of the most important characteristics of item response theory" (p. 35). Thus the evaluation of the extent to which item parameter invariance is realized in practice is itself an important concern. Of particular interest in this regard is the degree to which item parameters on achievement tests are invariant for groups with different instructional experiences.

It is intuitively reasonable that achievement test items should be sensitive to differences in instructional experiences. As stated by Masters (1988), "If the content of an item has been emphasized in an instructional program but either not taught or treated only superficially in another program, then that item is likely to be differentially difficult for students in those two instructional groups" (p. 18). In a similar vein, one might expect that those items that correspond to content emphasized in an instructional program would be differentially difficult before and after the instruction for the same group of students.

The Second International Mathematics Study (SIMS) provides a rich data source for investigating both of these possibilities. The opportunity to learn measures provide a means of

studying the differential effects of instructional exposure. Research reported by Miller and Linn (1988), Muthen (1988, 1989) and by Muthen, Kao, and Burstein (1988), indicates that differences in opportunity to learn can affect item parameters.

In addition to the opportunity to learn measures, the SIMS data for 8th grade students in the U.S. can be used to study the effects of differences in the types of math courses students take. The study identified four class types for 8th graders: remedial math classes, typical math classes, enriched math classes and algebra. The relatively difficulty of items has been found to vary as a function of class type.

Another feture of SIMS that makes it useful for investigaing instructional effects on items is its longitudinal design. Students in the eighth grade were administered a common core of items as a pretest in the fall and a posttest the following spring. Thus, the study also provides a basis for judging differential effects of instruction on the core math test items for a single group of students at two points in time.

A simple illustration of the latter type of comparison is presented in Figure 5. The letters in the figure correspond to the classification of the items by the content categories of measurement, arithmetic, geometry, and algebra. As can be seen, the ordering of most items in terms of difficulty is similar for most of the items when based on the results of the fall pretest or the spring posttest. There are two obvious outliers in the figure, however. Those two items are two of the seven items that were classified as algebra items by SIMS. The fact that they are below the main diagonal of the plot indicates that the items were relatively much easier in the spring than they were in the fall.

Since the students had received instruction in algebra during the interval between the pretest and the posttest it seems quite reasonable that algebra items should become relatively easier in the spring. Thus the question may not be so much why these two items appear as outliers as why the other five algebra items do not. To get some sense of this it may help to look at the specific items.

The seven algebra items are shown in Table 2 along with the simple proportion of students who answered each item correctly in the fall and in the spring. The first two items are the outliers, that is the ones that were much easier relatively in the fall than the spring. Though all seven items are classified as algebra items, it is evident that they vary a good deal both in the likelihood that students knew how to solve them before taking algebra and in the degree to which they are standard problems that students would be expected to learn how to solve in an eighth grade algebra class. Item number 3 is the only item that students are very unlikely to be able to answer correctly in the fall and which is a clear part of any reasonable instruction in first year algebra.

Item 16 is really a signed-number arithmetic problem rather than an algebra problem and presumably is classified with the algebra problems because it does not belong with the other three categories and because signed-number arithmetic is typically taught in eighth grade algebra classes. Although the remaining items obviously can be solved using algebraic principles can also be solved in other ways. Based on my little post hoc analyses, it seems reasonable the only items 3 and 16 are clear outliers.

It obviously would be unwise to generalize from this small example of instruction having a differential impact on the difficulty of 2 of 35 mathematics items. It illustrates three points that I want to make, however. First, effects may appear to be highly specific and limited to a relatively small fraction of the items. Second, and closely related, a general test of achievement may contain very few items that fall into a narrowly defined content category such as the multiplication of negative numbers, or even a somewhat broader category such as signed number arithmetic. Third, rather large differences in instructional experience may be needed to identify major effects. I think these three points are relevant not only to the small example in Figure 5, but to gaining an understanding of the results of a number of more substantial analyses of the differential impact of curricula and instruction on item parameters.

In is not hard to find examples such as the one in Figure 5. The important question, however, is not whether there are situations in which differences in instructional experiences lead to violations of IRT assumptions. Rather. the question is whether differences that are likely to be encountered in practice are sufficient to invalidate particular applications and interpretations. The results of several empirical studies that have been conducted in the past few years yield apparently conflicting results in this regard. Some studies suggest that differences in instructional experience can have substantial differential effects on some subsets of items, while other studies suggest that variations in curricula and instruction have little effect on tests or subsets of items.

Mehrens and Phillips (1986; 1987, also Phillips & Mehrens, 1987) have conducted several recent studies that have focused on traditional curricula and textbooks that are widely used throughout the country. Those studies suggest that textbook differences and differences in curricula that occur within school districts have little effect on the measurement characteristics of standardized tests. For example, Mehrens and Phillips (1987) found that, whether measured by ordinary p-values or Rasch item parameter estimates, there was a relatively close agreement between item difficulties on the Stanford Achievement Test for three groups of students using different textbooks.

Although the Mehrens and Phillips studies suggest that there is little reason to be concerned about differential instructional effects for the tests and textbooks they have investigated, some other studies suggest that the effects are sometimes more substantial. As was previously indicated, analyses of the SIMS data using the opportunity to learn measure Miller and Linn (1988), by Muthen (1989), and by Muthen, Kao, and Burstein (1988), suggest that differences in opportunity to learn can have important differential effects on item performance. A recent study reported by Masters (1988) indicated that the item response functions for a few items are quite different for groups of students who were enrolled in different high school math courses. Cook, Eignor, and Taft (1988) found that item parameters estimates on a biology test were quite sensitive to the recency with which students had taken a biology course. They concluded that a set of common items "clearly measured different attributes when given to a spring and fall sample and very similar attributes when given to two fall samples" (Cook, Eignor, & Taft, 1988, p. 43).

The conflict between results obtained by the latter authors and those that were obtained in studies conducted by Mehrens and Phillips may be more apparent than real. The series of studies conducted by Mehrens and Phillips have focused on reading and mathematics in the elementary grades and on widely used textbooks. On the other hand, the studies where more substantial differential effects have been found generally have focused on students at higher grade levels and have involved qualitatively greater differences in the nature of instructional experiences (e.g., different courses of study, or item and student specific ratings of opportunity to learn).

Also, even in studies with quite substantial differences in experiences, the effects are often limited to a relatively small fraction of the items in most of the studies. Muthen, Kao, and Burstein (1988), for example, found substantial sensitivity to exposure to instruction on only about 10 to 15% of the items in their analyses. Masters (1988) found only a slightly higher fraction of the items to be differentially affected by exposure to markedly different courses of instruction.

The implications of these results for the validity of IRT-based measures of achievement depend heavily on the use that is made of the information, the purposes of the tests, and the interpretations that are to be made of the scores. If items that are found to be most sensitive to instruction are eliminated so that the IRT assumptions are better satisfied, then there is a real danger that IRT will do more to decrease than to increase the validity of achievement test scores. Elimination of such items redefines the achievement domain in unknown ways and is likely to exacerbate the previously mentioned tendency to produce achievement tests that are overly saturated with aptitude. Such an outcome could have very negative consequences for

the educational uses of achievement tests. As Traub and Wolfe (1981) said several years ago: "It is bad enough that educators equate achievement with what can be captured in a test item. To limit our conception of achievement even further to only those items that fit a unidimensional latent trait model is to narrow our emphasis too much" (p. 383).

Elimination of items is, of course, only on way of responding to information regarding the lack of item invariance, or for that matter, the original item parameter estimates. There are other uses that can be made of the information, which can, in fact, enhance the validity of achievement test scores. Three such possibilities are briefly mentioned in closing.

First, and probably most important, the information can be useful in considerations of the construct validity of the achievement test scores and in decisions about what is being and what should be measured by the test. The Cook, Eignor, and Taft (1988) results illustrate the point. Their finding that the recency of instruction in biology had a marked effect on what was being measured by the biology achievement test has important implications for decisions about what such a test should measure and therefore for the content specifications of the test. As they suggest the results should force serious consideration of the importance of measuring "immediate end-of-course outcomes" versus "perhaps more enduring concepts" (p. 44).

A second potentially important use of information about the degree of item parameter invariance is illustrated by the recent work of Bock, Muraki, and Pfeiffenberger (1988) on item parameter drift. Differential change in item parameters over time poses a problem for the maintenance of a common scale that can be used for tracking achievement over time. However, as Bock, Muraki, and Pfeiffenberger have shown, differential drift was relatively steady, at least for the test they analyzed, and changes in item difficulties can be reasonably modeled by linear functions. Thus the differential drift can be taken into account. In addition, however, study of the content characteristics of items that show relatively steady increases in difficulty over a number of years, versus those that show relatively steady decreases in difficulty and those that have relatively constant difficulty can provide potentially useful information about the changing nature of the curriculum and of student achievement.

Finally, differential item sensitivity to instructional experiences could be used to help expand the nature of measured achievement domains in ways that will provide measures that are more useful for instruction. Rather than using such information to eliminate items, it can be used to identify content categories that may need to be expanded if we are to have achievement measures that are sensitive to differences in instruction. Together with more powerful analytical techniques such as those being developed by Muthen (1985, 1988) and by Tatsuoka (in press), expanded content categories may make it possible to understand instructional effects on achievement better and, possibly provide more diagnostically useful information about student achievement.

Conclusion

Has item response theory increased the validity of achievement test scores? As was indicated at the beginning, the question seems to defy a simple yes or no answer. Indeed, it probably is not even the right question. Better questions should deal with the ways in which IRT has and can contribute to increased validity and the ways in which it may decrease validity.

By raising fundamental questions about issues such as those illustrated by the controversy over scale properties or the content representation required for valid normative comparisons based on customized tests, IRT has increased the likelihood that more valid interpretations will be made of achievement test scores. It also has the potential of contributing to validity by forcing more careful consideration of content specifications and pointing to situations where better coverage of sparsely sampled content areas is needed. As is true of other technology, however, IRT can have negative consequences if misused. Assuming that content can be ignored on a customized test because the items have been calibrated is one clear example of such misuse. Limiting our definition of achievement to items that fit a

unidimensional IRT model for a relatively broad content domain such as mathematics concepts and applications or achievement in a subject area such as biology would be a more serious mistake, one that would damage rather than enhance the measurement of achievement.

References

Allen, N. L., Ansley, T. N., & Forsyth, R. A. (1987). The effect of deleting content-related items on IRT ability estimates. *Educational and Psychological Measurement, 47,* 1141-1152.

American Psychological Association, American Educational Research Association, National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anastasi, A. (1958). *Differential psychology* (3rd ed.). New York: Macmillan.

Anderson, R. C. (1972). How to construct achievement test to assess comprehension. *Review of Educational Research, 42,* 145-170.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bock, R. D., Mislevy, R., & Woodson, C. (1982). The next stage in educational assessment. *Educational Researcher, 11,* 4-11.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25,* 275-285.

Buros, O. K. (1948). Criticisms of commonly used methods of validating achievement test items. In *Proceedings of the 1948 ETS Conference on Testing Problems.* Princeton, NJ: Educational Testing Service.

Buros, O. K. (1977). Fifty years in testing: Some reminiscences, criticisms and suggestions. *Educational Researcher, 6,* 9-15.

Burket, G. (1984). Response to Hoover. *Educational Measurement: Issues and Practice, 3,* No. 4, 15-16.

Camilli, G. (1988). Scale shrinkage and the estimation of latent distribution parameters. *Journal of Educational Statistics, 13,* 227-241.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement, 25,* 31-45.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.), *New directions for testing and measurement: Measuring achievement, progress over a decade. Proceedings of the 1979 ETS Invitational Conference* (pp. 99-108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (Proceedings of a symposium in honor of Lloyd G. Humphreys) (pp. 147-171). Urbana, IL: University of Illinois Press.

CTB/McGraw-Hill. (1984). *Comprehensive Tests of Basic Skills, Forms U and V, Technical Report.* Monterey, CA: CTB/McGraw-Hill.

CTB/McGraw-Hill. (1986). *California Achievement Tests, Forms E and F. Norms Book, March through June.* Monterey, CA: CTB/McGraw-Hill.

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GEs. *Educational Measurement: Issues and Practice, 3*(4), 8-14.

Hoover, H. D. (1988). Growth expectations for low-achieving students: A reply to Yen. *Educational Measurement: Issues and Practice, 7*(4), 21-23.

Jones, L. V. (1988). Educational assessment as a promising area for psychometric research. *Applied Measurement in Education, 1*, 233-241.

Linn, R. L. (1985). Review of Comprehensive Tests of Basic Skills, Forms U and V. In J. V. Mitchell (Ed.), *The ninth mental measurements yearbook*. Lincoln, Nebraska: Buros Mental Measurements Institute, pp. 382-386.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25*, 15-29.

Mehrens, W. A. & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23*, 185-196.

Mehrens, W. A. & Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24*, 357-370.

Messick, S. (1988). Validity. In R. L. Linn (Ed.), *Educational measurement,* Third Edition (pp. 13-103). New York: Macmillan.

Miller, M. D. & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*, 205-219.

Mislevy, R. J. (1983). Item response models for grouped data. Journal of Educational Statistics, 8, 271-288.

Mislevy, R. J. (1988). Scaling procedures. In A. E. Beaton (Ed.), *Expanding the new design: The NAEP 1985-86 technical report* (pp. 177-204). Princeton, NJ: Educational Testing Service.

Muthen, B. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. Journal of Educational Statistics, 10, 121-132.

Muthen, B. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variable. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: Erlbaum.

Muthen, B. (1989). *Instructionally sensitive psychometrics: Applications to the Second International Mathematics Study.* Paper presented at Second International Mathematics Study Research Conference. Champaign, Illinois. January, 1989.

Muthen, B., Kao, C-F., & Burstein, L. (1988). *Instructional sensitivity in mathematics achievement test items: Application of a new IRT-based detection technique.* Paper presented at the annual meeting of the American Educational Research Association. New Orleans, April, 1988.

Pandey, T. N. & Carlson, D. (1983). Application of item response models to reporting assessment data. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 212-229). Vancouver, BC: Educational Research Institute of British Columbia.

Phillips, S. E. & Clarizio, H. F. (1988a). Limitations of standard scores in individual achievement testing. *Educational Measurement: Issues and Practice, 7*(1), 8-15.

Phillips, S. E. & Clarizio, H. F. (1988b). Conflicting growth expectations cannot both be real: A rejoinder to Yen. *Educational Measurement: Issues and Practice, 7*(4), 18-19.

Phillips, S. E. & Mehrens, W. A. (1987). Curricular differences and unidimensionality of achievement test data: An exploratory analysis. *Journal of Educational Measurement, 24*, 1-16.

Shepard, L. (1985). Review of Comprehensive Tests of Basic Skills, Forms U and V. In J. V. Mitchell (Ed.), *The ninth mental measurements yearbook*. Lincoln, Nebraska: Buros Mental Measurements Institute, pp. 386-389.

Snow, R. E. & Lohman, D. F. (1988). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement*, Third Edition (pp. 263-331). New York: Macmillan.

Tatsuoka, K. K. (in press). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. L. Glaser, A. M. Lesgold, & m. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition*. Hillsdale, NJ: Erlbaum.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Traub, R. E. & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. In D. C. Berliner (Ed.), *Review of research in education* (Vol. 9, pp. 377-435). Washington, DC: American Educational Research Association.

Wainer, H. & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12*, 339-368.

Way, W. D., Forsyth, R. A., & Ansley, T. N. (1989). IRT ability estimates from customized achievement tests without representative content sampling. *Applied Measurement in Education, 2*, 15-35.

Willingham, W. W. (1980). New methods and directions in achievement measurement. In W. B. Schrader (Ed.), *New directions for testing and measurement: Measuring achievement, progress over a decade. Proceedings of the 1979 ETS Invitational Conference* (pp. 73-80). San Francisco: Jossey-Bass.

Wright, B. D. (1968). Sample free test calibration and person measurement. *Proceedings of the 1967 ETS Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.

Yen, W. M. (1983). Use of the three-parameter model in the development of a standardized achievement test. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 123-141). Vancouver, BC: Educational Research Institute of British Columbia.

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50*, 399-410.

Yen, W. M. (1986). The choice of scale for educational measurement: an IRT perspective. *Journal of Educational Measurement, 23*, 2--235.

Yen, W. M. (1988). Normative growth expectations must be realistic: A response to Phillips and Clarizio, *Educational Measurement: Issues and Practice, 7*(4), 16-17.

Yen, W. M., Green, D. R., & Burket, G. R. (1987). Valid normative information from customized achievement tests. *Educational Measurement: Issues and Practice, 6*, 7-13.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24*, 293-308.

Appendix A

Figure 1

Plots of Scale Scores for Selected Percentiles by Grade on Four IRT Scaled Reading Tests

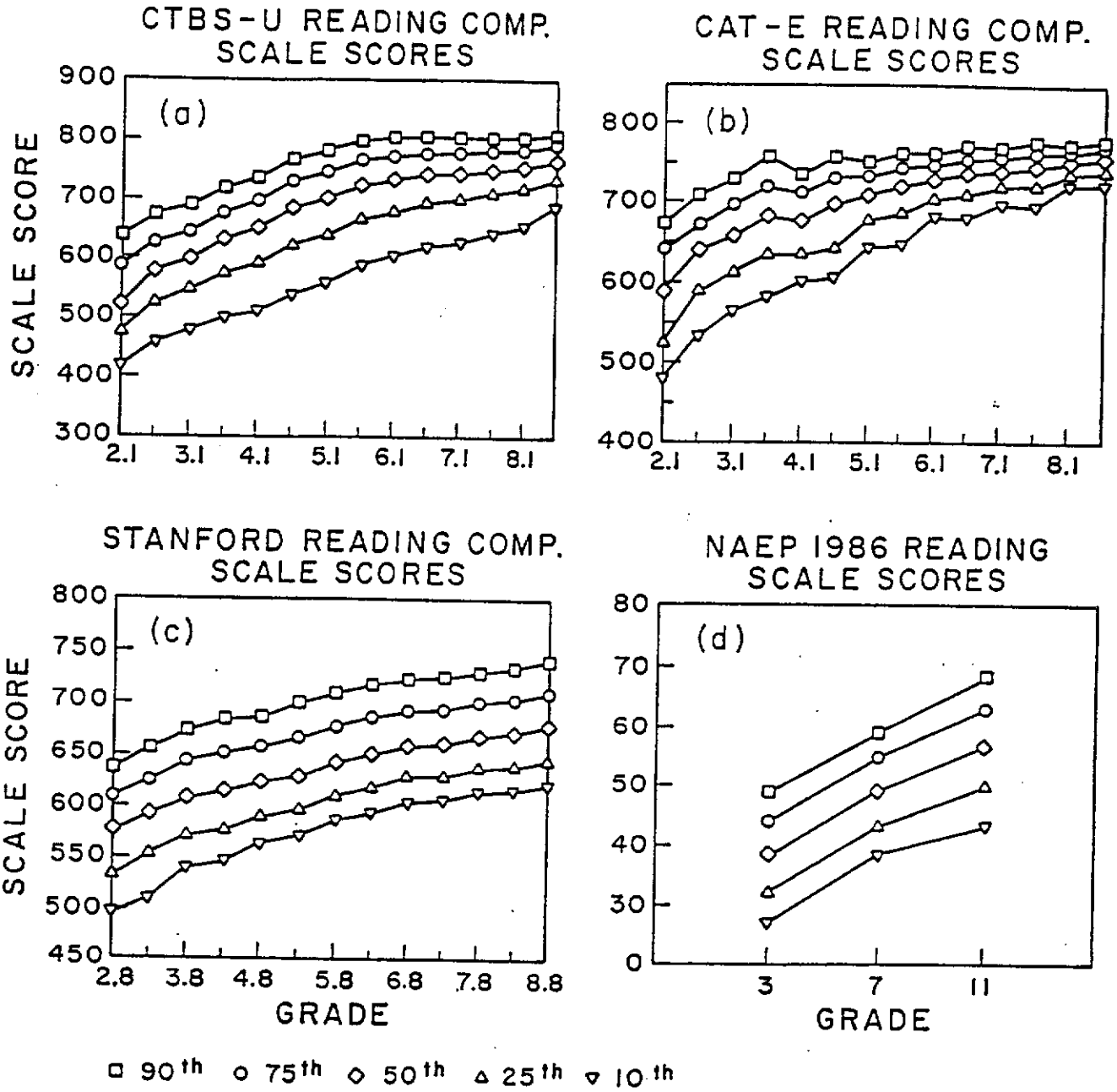


Figure 2

Plots of Scale Scores for Selected Percentiles by Grade on

Four IRT Scaled Mathematics Tests

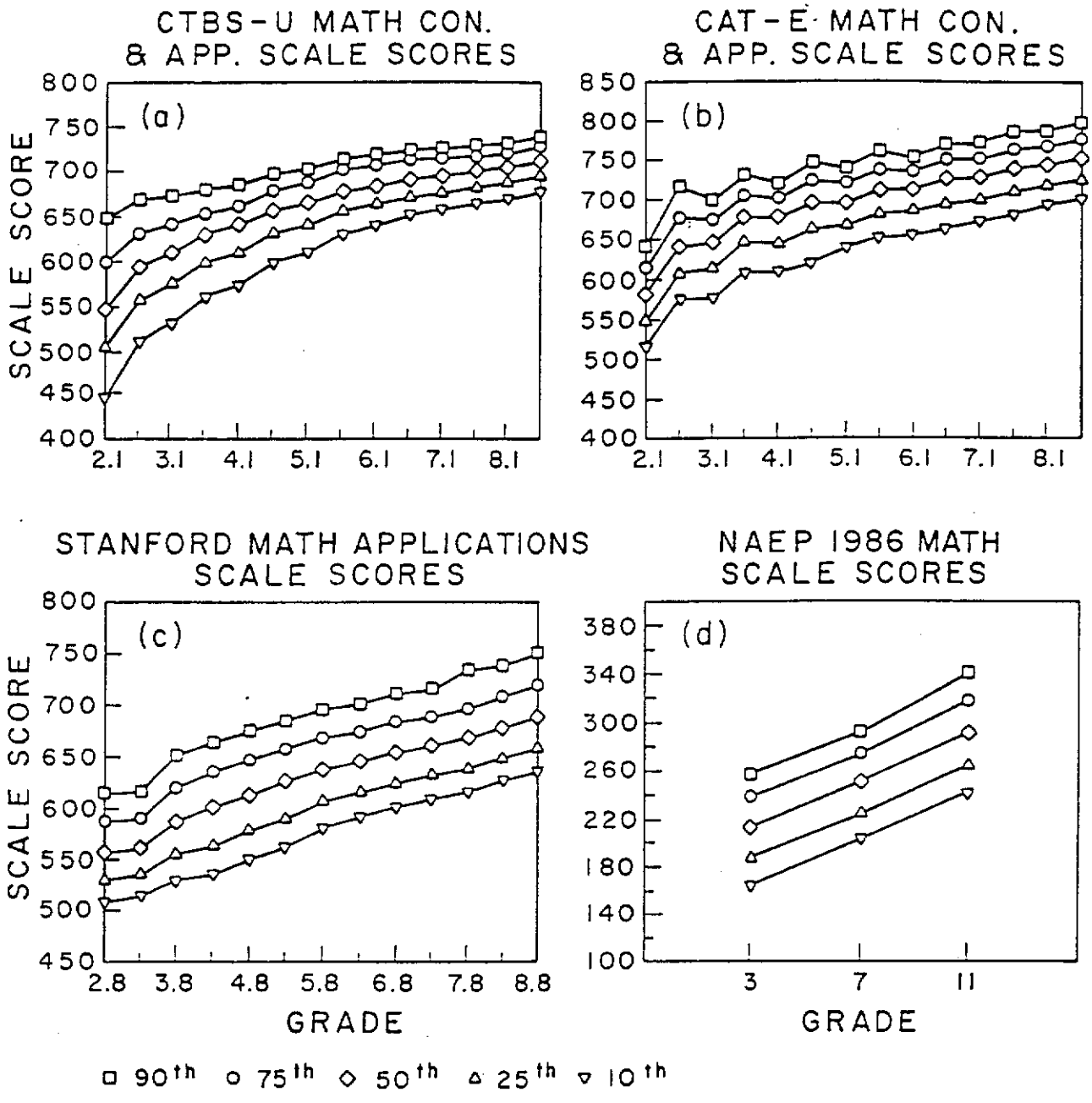


Figure 3

Fall and Spring Differences in Scaled Scores Between the 90th and 10th Percentiles on Two Mathematics Tests for Grades 2 through 8

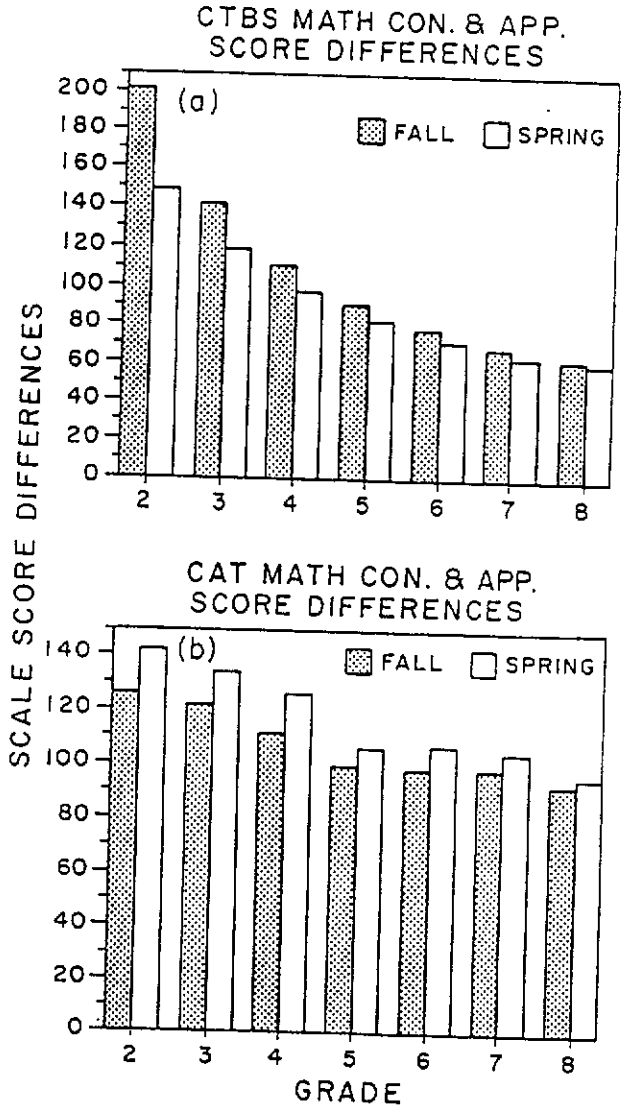


Figure 4

Distributions of Grade 5 Mathematics Test Scores for Four Years

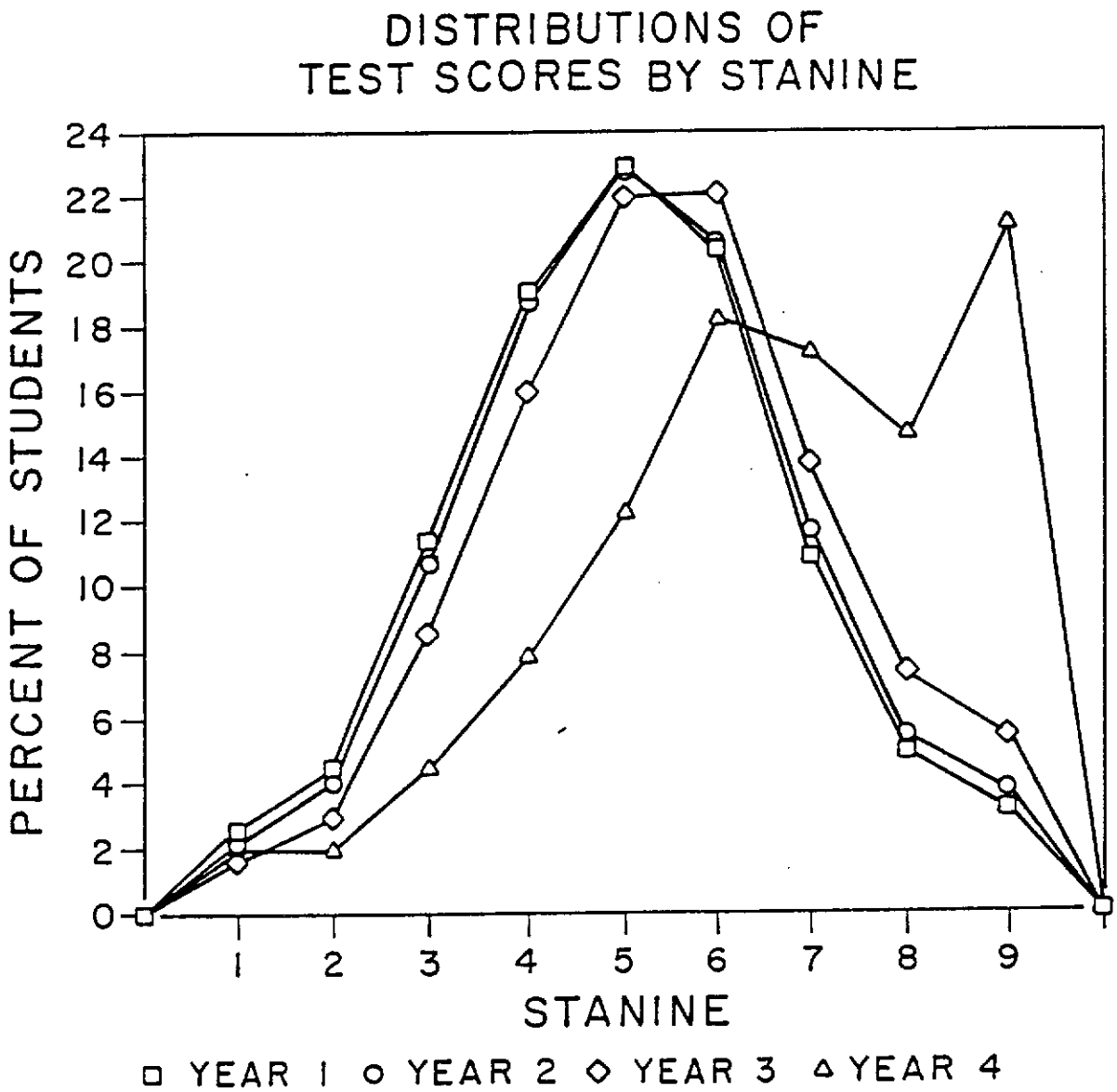
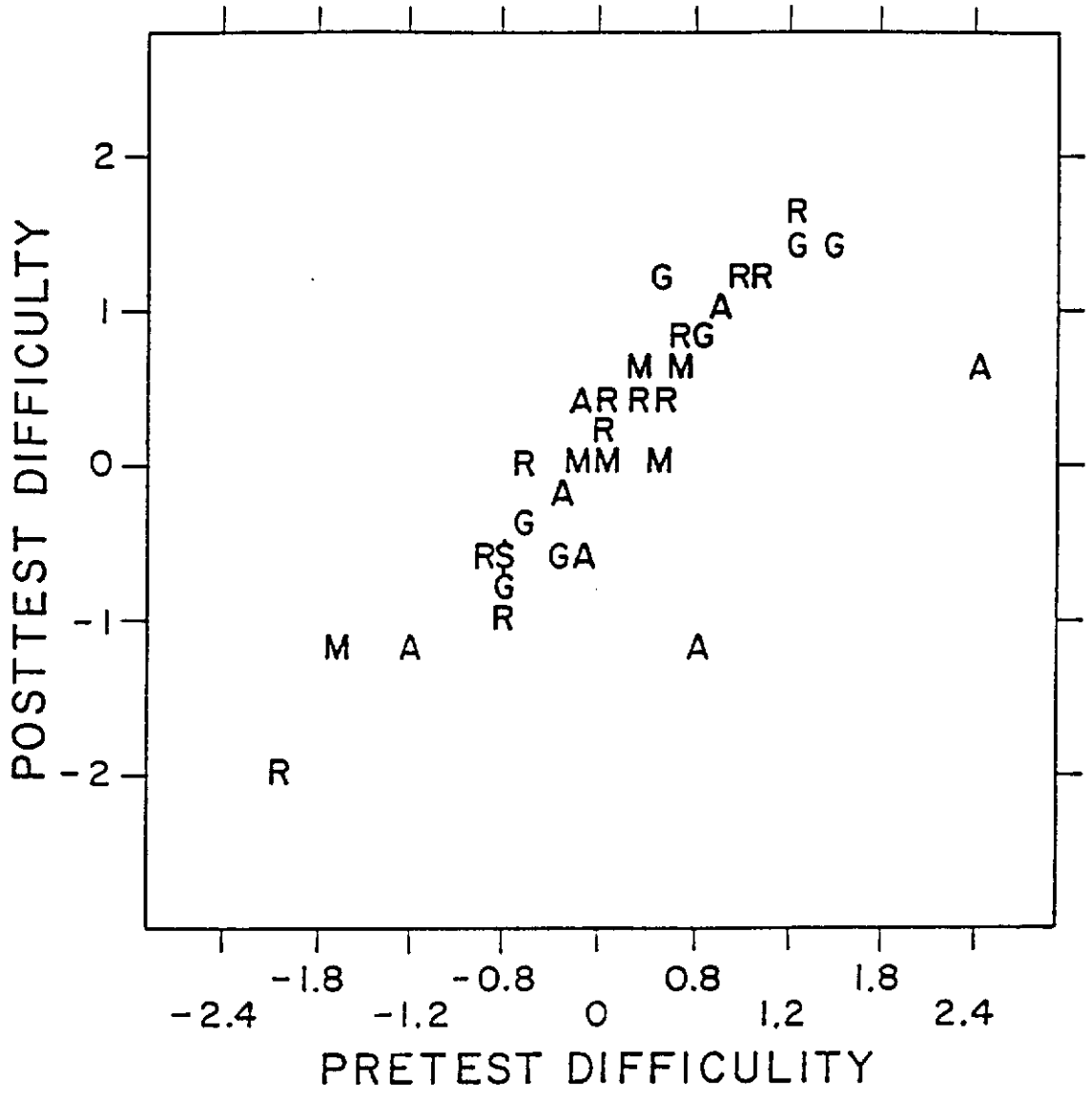


Figure 5

Scatterplot of Posttest Item Difficulties with Pretest Item Difficulties for Students Enrolled in Eighth Grade Algebra Classes (M = Measurement, R = Arithmetic, A = Algebra, G = Geometry, $ = Multiple Occurrence)



Appendix B

Table 1

Ratios of Standard Errors of Measurement at Scale Scores Corresponding to Selected Percentile Points to the Minimum Standard Error of Measurement

CTBS/U Math Concepts and Applications[1]

Percentile	Grade 2 (Level D) Fall	Grade 2 (Level D) Spring	Grade 3 (Level E) Fall	Grade 3 (Level E) Spring
95	2.1	3.1	2.9	3.0
90	1.6	2.2	1.8	2.0
75	1.0	1.2	1.2	1.3
50	1.2	1.0	1.1	1.1
25	1.7	1.2	1.6	1.1
10	4.4	1.6	4.1	2.0
5	8.0	2.9	7.4	3.6

CAT/E Math Concepts and Applications[2]

Percentile	Grade 2 (Level 12) Fall	Grade 2 (Level 12) Spring	Grade 3 (Level 13) Fall	Grade 3 (Level 13) Spring
95	1.3	4.2	1.1	1.4
90	1.1	2.6	1.1	1.4
75	1.1	1.5	1.0	1.1
50	1.1	1.1	1.0	1.0
25	1.2	1.0	1.0	1.0
10	1.6	1.1	1.2	1.1
5	2.1	1.2	1.5	1.2

1. Standard errors of meaurement estimated from Table 72 of CTBS Forms U and V Technical report using linear interpolation (CTB/McGraw-Hill, 1984).

2. Standard errors of meaurement estimated from Tables 5 and 6 of CAT Forms E and F Norms book using linear interpolation(CTB/McGraw-Hill, 1986).

Table 2

"Algebra items" classified by changes in difficulty from fall to spring for students in eighth grade algebra classes (Fp=fall proportion correct, Sp=spring proportion correct)

A. Items that are much easier in the spring than the fall.

Item #3: Fp=.16 Sp=.64
If $5x + 4 = 4x - 31$ then x is equal to
a. -35 b. -27 c. 3
d. 27 e. 35

Item #16: Fp=.53 Sp=.89
(-2) x (-3) is equal to
a. -6 b. -5 c. -1
d. 5 e. 6

B. Items that are easier in the spring than the fall.

Item #27: Fp=.69 Sp=.82
A shopkeeper has x kg of tea in stock. He sells 15 kg and then receives a new lot weighing 2y kg. What weight of tea does he now have?
a. x - 15 - 2y b. x + 15 + 2y c. x - 15 + 2y
d. x + 15 - 2y e. none of these

C. Items that are slightly easier in the spring than the fall.

Item #13: Fp=.87 Sp=.89
If P = LW and if P = 12 and L = 3, then W is equal to
a. 3/4 b. 3 c. 4
d. 12 e. 36

Item #18: Fp=.71 Sp=.78
If 4x/12 = 0, then x is equal to
a. 0 b. 3 c. 8
d. 12 e. 16

Item #30: Fp=.50 Sp=.57
The table below compares the height from which a ball is dropped (d) and the height it bounces (b).

d	50	80	100	150
b	25	40	50	75

Which formula describes this relationship?
a. $b = d^2$ b. b = 2d c. b = d/2
d. b = d + 25 e. b = d - 25

D. Items that are slightly harder in the spring than the fall.

Item #25: Fp=.69 Sp=.67
The air temperature at the foot of a mountain is 31 degrees. On top of the mountain the temperature is -7 degrees. How much warmer is the air at the foot of the mountain?
a. -38 degrees b. -24 degrees c. 7 degrees
d. 24 degrees e. 38 degrees