Psychological Assessment
1996, Vol. 8, No. 4, 341-349

1040-3590/96/$3.00

The New Rules of Measurement

Susan E. Embretson
University of Kansas

Classical test theory, which the authors maintain applied psychologists are still too often exclusively taught, are contrasted with the new rules of measurement. In the newer, model-based version of test theory, called item response theory (IRT), some well-known rules of measurement no longer apply. Six old rules of measurement that conflict with the new rules are reviewed, and intuitive explanations of the new rules are provided. Readers are also directed to additional informational sources about IRT, which, it is argued, every psychologist should be familiar with.

In an ever-changing world, psychological testing remains the flagship of applied psychology. Although both the context of application and the legal guidelines for using tests have changed, psychological tests themselves have been relatively stable. Many historically valued tests, in somewhat revised forms, remain in active current use. Further, although several new tests have developed in response to contemporary needs in applied psychology, the principles underlying test development have remained constant. Or have they?

Classical test theory has served test development well over several decades. Gulliksen's (1950) classic book, reprinted even in the 1990s, is often cited as the defining volume. However, classical test theory is much older. Many procedures were pioneered by Spearman (1907, 1913). Most psychologists should, and in fact do, know its principles. In some graduate programs, classical test theory is presented in a separate course that is required for applied psychologists and elective for other areas. In other graduate programs, classical test theory is part of the basic curriculum in testing methods for courses for clinical, counseling, industrial–organizational, and school psychologists.

However, since Lord and Novick's (1968) classic book introduced model-based measurement, a quiet revolution has occurred in test theory. Model-based measurement, known as *item response theory* (IRT) or *latent trait theory*, has rapidly become mainstream as a theoretical basis for psychological measurement. Increasingly, tests are developed from model-based measurement not only because the theory is more plausible but also because the potential to solve practical testing problems is greater.

A large family of diverse IRT models are now available to apply to an assortment of measurement tasks. IRT applications to available tests will be increasing. Although the early IRT models emphasized dichotomous item formats (e.g., the Rasch model), extensions to other item formats, such as rating scales (Andrich, 1982) and partial credit scoring (Masters, 1982) are now available. Further, the undimensional IRT models have been generalized to multidimensional models such that traits may be measured by comparisons within tasks (Kelderman & Rijkes, 1994), changes across conditions (Embretson, 1991), subtasks representing underlying cognitive components (Embretson, 1984), or conditioning on test-taking strategy (Rost, 1990). Also, the traits may be trait scores, personality traits, dispositions, or even attitudes.

To provide continuity between the new test theory and the old test theory, Lord and Novick (1968) derived many classical test theory principles from IRT. On the surface, this is good news to the busy applied psychologist who knows classical test theory but not IRT. The existence of derivations seemingly suggests that the rules of measurement, although rooted in a more sophisticated body of axioms, remain unchanged.

However, the new rules of measurement are fundamentally different from the old rules. Many old rules, in fact, must be revised, generalized, or even abandoned. This article has several goals. First, to illustrate the depth of the differences between measurement principles, a few well-regarded old rules will be compared with the corresponding new rules. Second, the basis of these new rules will be explained in nonstatistical language. Third, the reasons why the new rules of measurement are not widely known among psychologists will be explored.

A Comparison of Measurement Rules

Several old "rules" of measurement may be gleaned from the principles of classical test theory or its common extension. Other old "rules" are implicit in many applied test development procedures. Table 1 shows six old rules that will be reviewed here. The six old rules are followed by six corresponding new rules, which obviously conflict with the old rules.

I would argue that the old "rules" represent common knowledge or practice among psychologists. These rules have guided the development of many, but certainly not all, published psychological tests. Obvious exceptions are tests that are developed by large-scale testing corporations such as the Educational Testing Service (ETS) and the American College Testing Program (ACT), in which non-IRT procedures have been developed to circumvent the limitations of some old rules. That is, nonlinear test equating (see Holland & Rubin, 1982) and population-free item indexes, such as the delta index used by ETS (see Gullik-

An earlier version of this article was presented at the 103rd Annual Convention of the American Psychological Association, New York, August 1995.

Correspondence concerning this article should be addressed to Susan E. Embretson, Department of Psychology, 426 Fraser Hall, University of Kansas, Lawrence, Kansas 66045.

Table 1
Old and New Rules of Measurement

Rule no.	Old rule	New rule
1.	The standard error of measurement applies to all scores in a particular population.	The standard error of measurement differs across scores, but generalizes across populations.
2.	Longer tests are more reliable than shorter tests.	Shorter tests can be more reliable than longer tests.
3.	Comparing test scores across multiple forms depends on test parallelism or adequate equating.	Comparing scores from multiple forms is optimal when test difficulty levels vary across persons.
4.	Unbiased assessment of item properties depends on representative samples from the population.	Unbiased estimates of item properties may be obtained from unrepresentative samples.
5.	Meaningful scale scores are obtained by comparisons of position in a score distribution.	Meaningful scale scores are obtained by comparisons of distances from various items.
6.	Interval scale properties are achieved by selecting items that yield normal raw score distributions.	Interval scale properties are achieved by justifiable measurement models, not score distributions.

sen, 1950, p. 368), were developed to counter Old Rule 3 and Old Rule 4, respectively. However, these techniques are not well known outside large-scale testing programs, hence they are not routinely applied in the development of psychological tests. Thus, the old rules characterize substantial practice in test development.

Rule 1: The Standard Error of Measurement

Old rule 1. The standard error of measurement applies to all scores in a particular population.

New rule 1. The standard error of measurement differs across scores (or response patterns), but generalizes across populations.

Score interpretations depend on the standard error of measurement. The standard error of measurement is routinely used to construct confidence intervals for individual scores. The confidence intervals can guide score interpretations in several ways; for example, an individual's performance may be presented as a likely range of scores rather than a single score. The standard error of measurement also is routinely used to interpret differences between scores on different tests or subtests. It is important to note that a necessary assumption for constructing confidence intervals in classical test theory is that measurement error is distributed normally and equally for all score levels.

In classical test theory, the standard error of measurement is derived from population-specific estimates. The following well-known formula for the standard error of measurement involves an estimate of reliability, r_{tt}, and an estimate of variance, σ^2, as follows:

$$SE_{msmt} = (1 - r_{tt})^{1/2}\,\sigma. \qquad (1)$$

Because populations often differ in reliability and variability, Equation 1 will yield different standard errors of measurement.

Figure 1 shows the classical test theory (CTT) results for SE_{msmt} for two different populations on a 30-item test with a normal item difficulty range. The data in Figure 1 are based on 1,000 cases per population. SE_{msmt} is plotted by z scores representing trait levels. Consistent with the assumptions of CTT, SE_{msmt} is constant across trait levels but differs between the populations.

New Rule 1, from IRT, conflicts with both aspects of Old Rule 1. To show this concretely, a trait-level score and a corresponding SE_{msmt} were estimated using the Rasch (1960) IRT model for each person in the same two populations as for CTT. The Rasch model is most similar to CTT applications because the ability parameter is estimated from raw total scores. With the Rasch model, trait levels were estimated separately for each score or response pattern, controlling for the characteristics (e.g., difficulty) of the items that were administered.

The IRT values for SE_{msmt} in Figure 1 are identical for the two populations (noted as "IRT-All Populations"). In the max-

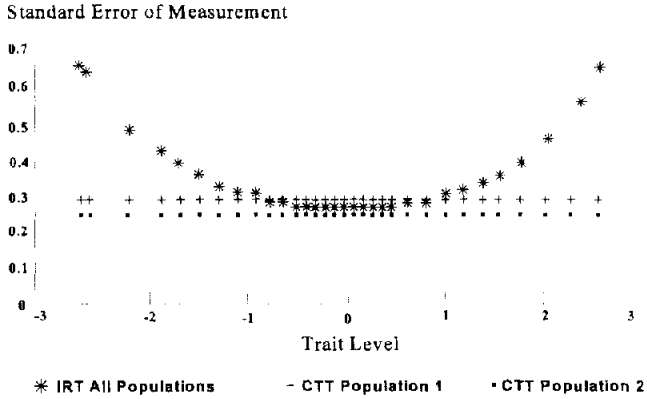


Figure 1. Two versions of measurement error. IRT = item response theory; CTT = classical test theory.

imum-likelihood estimation method, trait level and the corresponding SE_{msmt} estimates depend only on the total score or response pattern. However, Figure 1 shows that SE_{msmt} is not constant across trait levels. SE_{msmt} is lowest for moderate trait levels (i.e., z scores near zero) and highest for extreme trait levels. The difference in SE_{msmt} by trait level reflects the distribution of item difficulty. For extreme scores, tests usually contain too few items that are appropriate for the extreme trait level. Consequently, IRT estimates SE_{msmt} to be high for extreme scores.

In IRT measurement errors for each individual trait level are estimated, but it is also possible, as in CTT, to have a single value to describe the population. In IRT, the composite value is a mean of the individual values, whereas in CTT the single value applies to all trait levels. A composite value for SE_{msmt} for a population can be computed by averaging the IRT estimates across individuals. Thus, in Figure 1, the various SE_{msmt}'s would be weighted by the frequency of the abilities to which they correspond. For the IRT values in Figure 1, the composite SE_{msmt} is .33.

Rule 2: Test Length and Reliability

Old rule 2. Longer tests are more reliable than shorter tests.

New rule 2. Shorter tests can be more reliable than longer tests.

In CTT, the Spearman-Brown prophesy formula directly implies that longer tests are more reliable than shorter tests. Guilford (1954) showed that lengthening a test by a factor of n parallel parts results in true variance increasing more rapidly than error variance. If r_{tt} is the reliability of the original test, the reliability of the lengthened test r_{nn} may be anticipated as follows:

$$r_{nn} = \frac{nr_{tt}}{1 + (n-1)r_{tt}}, \qquad (2)$$

where n is the ratio of the number of new items to the number of old items. Equation 1 may also be applied to shortened tests. That is, if a test with a reliability of .86 is shortened to two thirds of the original length ($n = .667$), then the anticipated reliability of the shorter test is .80. Thus, in CTT, shorter tests generally imply increased measurement error. The new rule from IRT asserts that short tests can be more reliable than longer tests. In Figure 2, the IRT SE_{msmt} is plotted by trait level for four tests for 3,000 individuals. Two tests have fixed item content; that is, the same items are administered to each individual. In this case, it can be seen that IRT, like CTT, has lower SE_{msmt} error for the longer test. The composite SE_{msmt}'s for the 30-item and the 20-item tests, respectively, are .349 and .419.

But, notice that SE_{msmt} for another 20-item test plotted in Figure 2, namely an adaptive test, is drastically *smaller* for most trait levels than for the 30 item fixed content test. Adaptive test items are individually selected for a person to be optimally appropriate for his or her trait level. Thus, items that are too extreme for the person are avoided. A direct result of adaptive testing is to provide very small SE_{msmt} at all trait levels. The composite SE_{msmt} for the adaptive test in Figure 2 is .279, which is smaller than the corresponding composite for the 30-item test. The adaptive test has the same item discriminations as the items on the fixed content test.

Thus, the new rule of test length represents the advantage of adaptive tests or fixed-content tests. Notice that a longer adaptive test yields smaller SE_{msmt} error, like CTT. But, as shown by the composite measurement errors across trait levels, the shorter (adaptive) test can be more reliable than the longer normal-range test.

In fairness to classical test theory, it should be noted that an assumption underlying the Spearman-Brown prophesy formula is that the test is lengthened with parallel parts. An adaptive test, by its nature, does not to meet this assumption. However, the point here is that the old rule about test length and reliability conflicts sharply with current practice in adaptive testing, which is based on the new rule from IRT.

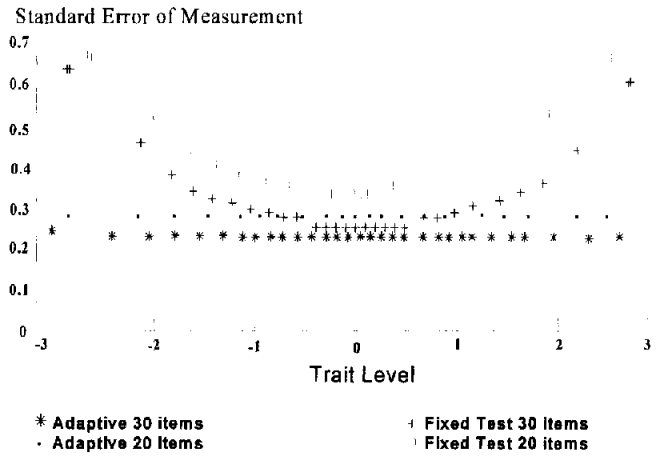


Figure 2. Measurement error and test length.

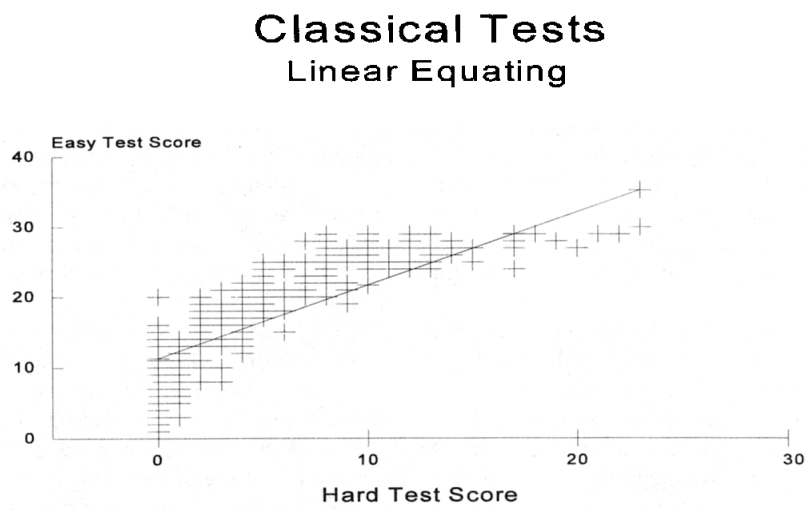
Rule 3: Interchangeable Test Forms

Old rule 3. Comparing test scores across multiple forms depends on test parallelism or test equating.

New rule 3. Comparing test scores across multiple forms is optimal when test difficulty levels vary between persons.

When individuals receive different test forms, some type of equating is needed to compare their scores. In traditional CTT, equating meant establishing that the different test forms were essentially equal. Gulliksen's (1950) classic text defined strict conditions for test parallelism in CTT, which included the equality of means, variances, and covariances across test forms. According to Gulliksen (1950), if the two tests meet the statistical conditions for parallelism, then scores may be regarded as comparable across forms.

More recent extensions of CTT have considered the test form equating issue more liberally, as score equivalencies between forms. So, for example, if an individual receives a score of 21 on Test Form A, an expected score on Test Form B may be given. Several procedures have been developed equating tests with different item properties, such as linear equating and equipercentile equating. These methods are used in conjunction with various empirical designs such as random groups or common anchor items (see Angoff, 1982, for a summary of several such methods). For a simplified example, suppose that both test



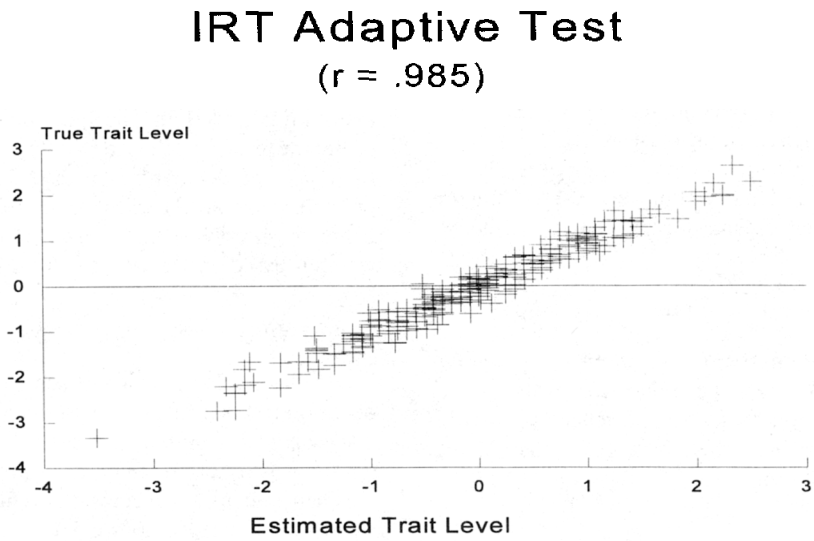


Figure 3. Relationship between scores on tests of different difficulties.

forms could be given to the same group with no carry-over effects. A very simple linear equating would involve regressing scores from one test form to the other test form. Score equivalencies between the test forms are established by using the regression equation to predict scores. This type of equating method can be applied to test forms that have different means, variances, and even reliabilities.

Although even wholly different tests can be linked by the newer equating methods, equating error can be problematic. That is, the score equivalencies represent expected values and individual fluctuations may be rather large. Equating error is influenced by differences between the test forms, especially in test difficulty level (see Peterson, Marco, & Stewart, 1982). Thus, test forms with high reliabilities and similar score distributions will be most adequately equated.

The effect of test difficulty differences on equating error may be readily seen in Figure 3, top panel. Data for 3,000 examinees were simulated for two 30-item test forms with equal item properties except for item difficulty level. A linear equating between the two test forms is given by the regression line from the easy to the hard test. Equating error is readily discerned on Figure 3 by the dispersion of scores around the regression line. For a hard test score of zero, examinees are observed with scores ranging from 0 to 20 on the easy test. The hard test simply does not have the floor to distinguish between these examinees, and so equating is not very satisfactory. Further, a *linear* equating is not

adequate to establish score correspondence. The relationship between test scores is obviously nonlinear. The upper panel of Figure 3 shows that easy test scores are underestimated at some score levels, overestimated at others. The low ceiling of the easy test compresses high scores, whereas the high floor of the hard test compresses low scores. Thus, a nonlinear regression is needed to fully describe score correspondence. The correlation between test scores is .805, which yields an expected true score correlation for either test of .897 (i.e., $.805^{1/2}$).

The IRT version of "equating" follows directly from the IRT model, which implicitly controls for item differences between test forms. The lower panel of Figure 3 shows the same simulation sample with trait level estimates obtained from IRT-scaled adaptive tests. Because adaptive test items are individually selected for a person, in large populations, hundreds of different "test forms" may be administered. For the data in the lower panel of Figure 3, item difficulty differences between "test forms" were directly controlled in the trait level estimates from the Rasch IRT model. Because the data are simulated, true trait level was known. So, this panel shows the regression of true trait level on the estimated trait scores from the adaptive tests. It can be seen that prediction is more accurate than in the top panel of Figure 3, as the correlation is .985.

Most important, better estimation of trait levels for all individuals are obtained from administering *different* test forms. More accurate estimation of each individual means that score differences are more reliable. Thus, the new rule from IRT means that *nonparallel* test forms that differ substantially, and deliberately, in difficulty level from other forms, yield better score comparisons.

Rule 4: Unbiased Assessment of Item Properties

Old rule 4. Unbiased assessment of item properties depends on representative samples from the target population.

New rule 4. Unbiased estimates of item properties may be obtained from nonrepresentative samples.

Assessing the classical item statistics of item difficulty (i.e., p values as the proportion passing) and item–total correlations (biserial correlations) yields noncomparable results if obtained from unrepresentative samples. Suppose that two biased samples are taken from a population: a low group, with scores below the mean, and a high group, with scores above the mean. These results are taken from the same simulation study as just described, such that approximately half the 3,000 cases fall in each group.

In the upper panel of Figure 4, the estimated p values for items are plotted from the two groups. A linear regression would indicate that the relative intervals between items is maintained. However, notice that the relationship between p values, although monotonic, is not linear. The distances between items with high p values is greater in the low group, whereas the distances between items with low p values is greater in the high group. The correlation between p values is only .800. The biserial correlations (not shown) of items with total score differs even greater between groups because the relationship was curvilinear.

The lower panel of Figure 4 shows the item difficulty values that are obtained by a Rasch model scaling of the same data as shown in Figure 4's top panel. In the lower panel, unlike in the top panel, the correspondence of item difficulty values is quite close between the two extreme groups. The correlation between item difficulty values in the lower panel is .997.

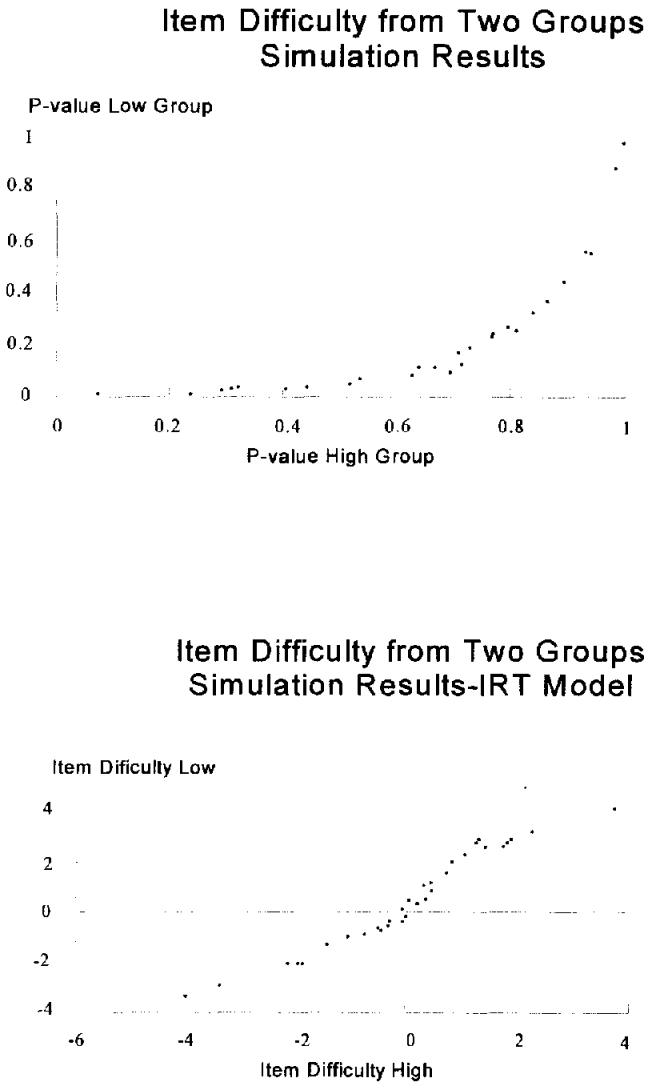


Figure 4. Relationship between item difficulties obtained from two groups.

Rule 5: Establishing Meaningful Scale Scores

Old rule 5. Meaningful scale scores are obtained by standard scores.

New rule 5. Meaningful scale scores are obtained from IRT trait score estimates.

Embretson and DeBoeck (1994) noted that test score meaning depends on specifying an appropriate comparison. A comparison is defined by two features: (a) the standard with which a score is compared and (b) the numerical basis of the comparison (order, difference, ratio, etc.).

In CTT, score meaning is determined by a norm-referenced

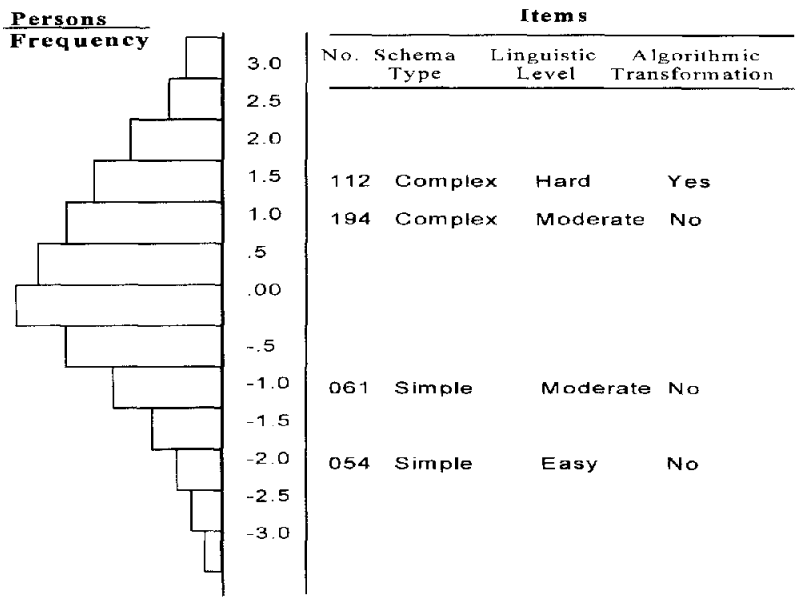


Figure 5. Common scale measurement of item difficulty and trait scores.

standard, and the numerical basis is order. That is, scores have meaning when they are compared with a relevant group of people for relative position. To facilitate this comparison, raw scores are linearly transformed into standard scores that have more direct meaning for relative position. An objection that is often raised to norm-referenced meaning is that scores have no meaning for what the person actually can do.

In IRT, a score is compared with items; persons and items are calibrated on a common scale. For example, on the right side of Figure 5 is a distribution of item difficulties. On the left side of the figure is the distribution of scores for a group of people. The match between trait level and item difficulty has direct meaning for expected item performance. For Item 112 (e.g., with a scale value of 1.5), people in the distribution who fall below this item are more likely to fail than to pass. That is, the probability that a person passes a particular item is derived from the match of item difficulty to trait level. As in psychophysics, an item is at the person's threshold when the person is as likely to pass as to fail the item. When an item's difficulty equals the person's trait level (i.e., in the Rasch model), then the person's probability of failing equals the probability of passing. Or, stated another way, the odds are 50/50 for passing versus failing. Thus, analogous to psychophysics, the item falls at the person's threshold. If the person's trait level exceeds the item, then the person is more likely to pass the item. Conversely, if a person's trait level is lower than item difficulty, then the odds are more favorable for failing the item.

To summarize, in IRT models, the meaning of a score can be referenced directly to the items. If these items are further structured by content, substantive trait level meaning can be derived. The sample items in the figure are from a mathematics word problem test. The substantive features of the sample items are shown on the right side of Figure 5. These features were derived from a cognitive model of mathematical problem solving. People who can solve Item 112 correctly are likely to be able to process a complex schema on linguistically difficult problems that require algorithmic transformations. Further details are given in Embretson (1995).

In some tests, particularly achievement tests, IRT trait levels are also linked to norms. In this case, IRT scores are linearly transformed to standard scores. Thus, IRT trait levels also can have norm-referenced meaning.

Rule 6: Establishing Scale Properties

Old rule 6. (Implicit) Interval scale properties of measures are achieved by selecting items to achieve normal raw score distributions.

New rule 6. Interval scale properties are achieved by justifiable measurement models.

Routine test development procedures for many psychological tests include selecting items to yield normal distributions in a target population. Even if normal distributions are not achieved in the original raw score metric, scores may be transformed or normalized to yield a normal distribution. These transformations are nonlinear, and so change the relative distances between scores.

Score distributions have implications for the level of measurement that is achieved. The hard test and easy test data in the top panel of Figure 3 yield skewed score distributions. In that panel, the best-fit linear regression is obviously inadequate to describe the nonlinear relationship between the test scores. Figure 6 uses the same data as Figure 3 to show how interval scale values are not achieved by CTT but are achieved by IRT. In the top panel of Figure 6, the values for the hard-versus-easy tests are shown again, here with the appropriate nonlinear regression curve.

To understand how score distributions on the two tests influence the level of measurement, consider an example of four persons, shown on Figure 6. On the hard test, Person 2's score is 5

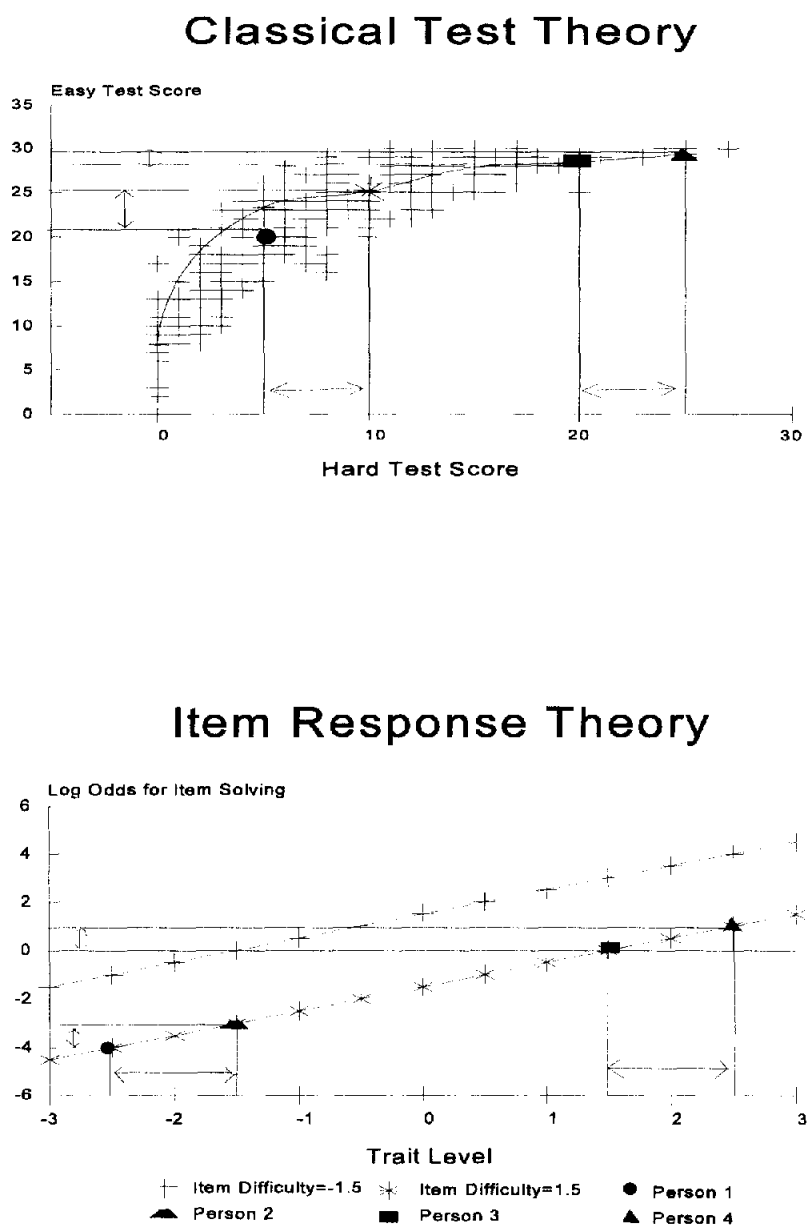


Figure 6. Relative distances of scores in classical test theory (CTT) and item response theory (IRT).

points higher than Person 1's score. The anticipated score difference on the easy test is approximately 4 points. Similarly, on the hard test, Person 4's score is 5 points higher than Person 3's score. However, on the easy test, their anticipated score differences are less than 1 point. Thus, the relative distances between scores are not the same from the hard to the easy test. Only ordinal level measurement has been achieved on one or both tests.

Jones (1971) pointed out that the classical methods to develop normally distributed trait scale scores will achieve interval scale measurement under certain assumptions. Specifically, these assumptions are that true scores (a) have interval scale properties and (b) are normally distributed in the population. Only linear transformations preserve score intervals as well as distribution shapes (see Davison & Sharma, 1990, for the latter). Thus, if raw scores are normally distributed, then only a linear transformation, such as a standard score conversion, will preserve score intervals to appropriately estimate true score. Notice, however, that scale properties are tied to a specific population. If the test is applied to a person from another population, can the interval scale properties still be justified? If not, then scale properties are population-specific.

For IRT models, particularly the Rasch model, several articles (e.g., Fischer, 1995; Roskam & Jansen, 1984) show how interval or even ratio scale properties are achieved. The Rasch model also has been linked to fundamental measurement (see

Andrich, 1988) because of the simple additivity of the parameters. A basic tenant of fundamental measurement is additive decomposition (see Michel, 1990), in that two parameters are additively related to a third variable. In the Rasch model additive decomposition is achieved; the log odds that a person endorses or solves an item is the simple difference between his or her trait level, θ_j, and the item's difficulty, b_i, as follows:

$$\text{LogOdd}_{ij} = \theta_j - b_i. \quad (3)$$

According to the principle of additive decomposition, interval scale properties hold if the laws of numbers apply. Specifically, the same performance differences must be observed when trait scores have the same interscore distances, regardless of their overall positions on the trait score continuum. Suppose, for example, that the trait level distance between Person 1 and Person 2 equals the distance between Person 3 and Person 4. Now, these intervals are justifiable if the same performance differential exists between the two pairs of persons. This property was *not* achieved for the CTT tests in the lower panel of Figure 6.

However, the relative performance differentials are maintained across tests of different difficulty for IRT. The lower panel of Figure 6 shows the regression of log odds for an easy ($b_i = -1.5$) and a hard item ($b_i = 1.5$) on trait level. The regression lines were obtained directly by applying Equation 3 to the various trait levels to each item. Also shown in the lower panel of Figure 6 are trait levels for four persons. As for the hard test in the upper panel of Figure 6, the trait level distance between Person 1 and Person 2 equals the trait level distance between Person 3 and Person 4. These relative distances are maintained in the log odds for both the easy and the hard item. Thus, regardless of item difficulty level, it can be seen that the same relative difference in performance holds. Maintaining relative distances over varying difficulties of items implies that a quality of interval scale measurement has been obtained.

Understanding and Diffusing the New Rules in Psychology

Model-based measurement is complex. Thus, a comprehensible explanation of the new rules is beyond the scope of this short article. A recent chapter (Embretson, in press-a) explains several new rules. Further, several books and edited volumes are available on IRT. Hambleton, Swaminathan, and Rogers's (1991) book is quite readable for a psychological audience, although it is limited somewhat by the examples, which were designed for readers in the education field.

Psychologists are not generally familiar with model-based measurement even though it is not new. Item response theory is often traced to Lord's (1952) monograph on item regression or to Rasch's (1960) book. Further, model-based measurement is increasingly part of test development practice. IRT scaling is now part of several major tests, including the computer adaptive form of the Armed Services Vocational Aptitude Battery (Department of Defense, 1996), the computerized form of the Scholastic Aptitude Test, and the latest revision of the Stanford-Binet (Thorndike, Hagen, & Sattler, 1986). As experts in testing, psychologists should know the fundamentals behind the development of these tests, namely, IRT.

Psychologists have had little exposure to model-based measurement for three reasons. First, IRT is statistically sophisticated as compared with classical test theory. A course on IRT is probably best preceded by a full-year course on graduate statistics. Further, a full course devoted only to IRT may be necessary to develop an adequate understanding. A chapter in a book on testing is most certainly insufficient. Second, and obviously interactive with the first reason, measurement courses have been declining in graduate schools in psychology. Aiken, West, Sechrest, and Reno's (1990) survey of graduate programs found that emphasis on measurement had substantially declined over the last 20 years. Graduate students are now barely exposed to classical test theory. So, the need for a full course on IRT has not fit in with the trends Aiken et al. (1990) observed. Third, as of this writing, no adequate textbook for psychologists exists on IRT. To be understood by psychology graduate students (and their professors), IRT must be connected to psychological concepts and illustrated by psychological data.

The textbook issue, obviously, must be addressed by someone in the small group of psychologists who are IRT experts. Currently, there is no such book available, probably primarily because IRT experts have been more concerned with technical issues than with expositional issues. However, the other two issues involve policy in graduate curriculum. Those psychologists who have a stake in the methodology surrounding testing need to plan for model-based measurement in the curriculum. If not, as model-based measurement becomes increasingly routine in testing, it is very possible that testing experts will no longer understand the principles underlying testing.

References

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology and measurement in psychology. *American Psychologist, 45,* 721–734.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR 20 index and the Guttman scale response pattern. *Educational Research and Perspectives, 9,* 95–104.

Andrich, D. (1988). *Rasch models for measurement.* Newbury Park, CA: Sage.

Angoff, W. (1982). Summary and derivation of equating methods used at ETS. In P. Holland & D. Rubin (Eds.), *Test equating.* New York: Academic Press (pp. 55–79).

Davison, M., & Sharma, A. R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regression. *Psychological Bulletin, 107,* 394–400.

Department of Defense. (1996). Computerized adaptive test (CAT) for the Armed Services Vocational Aptitude Test. Washington, DC: Author.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika, 49,* 175–186.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–516.

Embretson, S. E. (1995, August). The new rules of measurement. Paper presented at the annual meeting of the American Psychological Association, New York.

Embretson, S. E. (in press-a). Measurement principles for the new generation of tests: A quiet revolution. In R. Dillon (Ed.), *Handbook on testing.* Westport, CT: Greenwood.

Embretson, S. E. (in press-b). Multicomponent latent trait models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer-Verlag.

Embretson, S. E., & DeBoeck, P. (1994). Latent trait theory. In R. J. Sternberg (Ed.), *Encyclopedia of intelligence* (pp. 4021–4017), New York: Macmillan.

Fischer, G. (1995). Derivations of the Rasch model. In I. G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Holland, P., & Rubin, D. (1982). *Test equating*. New York: Academic Press.

Jones, L. V. (1971). The nature of measurement. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 485–498). Washington, DC: American Council on Education.

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*, 149–176.

Lord, F. (1952). A theory of test scores, *Psychometric Monographs* (whole No. 1).

Lord, F., & Novick, M. (1968). *Statistical theories of mental tests*. New York: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Michel, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Erlbaum.

Peterson, N., Marco, G., & Steward, E. (1982). A test of the adequacy of linear score equating models. In P. Holland & D. Rubin (Eds.), *Test equating* (pp. 71–135). New York: Academic Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute of Educational Research. (Expanded ed., 1980. Chicago: The University of Chicago Press)

Roskam, E., & Jansen, P. G. W. (1984). A new derivation of the Rasch model. M. E. Degreef & J. von Buggenhaut (Eds.), *Trends in mathematical psychology* (pp. 293–307). Amsterdam: Elsevier Science Publishers.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 3*, 271–282.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 15*, 201–292.

Spearman, C. (1913). Correlation of sums and differences. *British Journal of Psychology, 5*, 417–426.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Technical manual: Stanford–Binet Intelligence Scale* (4th ed.). Chicago: Riverside.

Received May 31, 1996
Revision received July 24, 1996
Accepted July 24, 1996 ■