On the Reliability of Testlet-Based Tests
Author(s): Stephen G. Sireci, David Thissen and Howard Wainer
Source: *Journal of Educational Measurement*, Vol. 28, No. 3 (Autumn, 1991), pp. 237-247
Published by: National Council on Measurement in Education
Stable URL: https://www.jstor.org/stable/1434845
Accessed: 25-02-2019 20:45 UTC

Journal of Educational Measurement
Fall 1991, Vol. 28, No. 3, pp. 237–247

On the Reliability of Testlet-Based Tests

Stephen G. Sireci
Fordham University
David Thissen
University of North Carolina
Howard Wainer
Educational Testing Service

If a test is constructed of testlets, one must take into account the within-testlet structure in the calculation of test statistics. Failing to do so may yield serious biases in the estimation of such statistics as reliability. We demonstrate how to calculate the reliability of a testlet-based test. We show that traditional reliabilities calculated on two reading comprehension tests constructed of four testlets are substantial overestimates.

Modern testing is beginning to move away from thinking that focuses on items as the molecular construct out of which tests are built. Test builders are beginning to use larger tasks as the fundamental units of which tests are manufactured. The psychometric foundation for such an idea was provided by Wainer and Kiely (1987) and expanded by Wainer and Lewis (1990); they called this new construct the *testlet.*

Thissen, Steinberg, and Mooney (1989) used the testlet approach in the analysis of a traditional reading comprehension test. In their example, there were four passages with 22 associated questions (7, 4, 3, and 8 items, respectively). They found that, when the testlet structure was explicitly characterized with a polytomous item response theory (IRT) model, it revealed that more traditional approaches overestimated the precision of measurement as characterized by the test information function. This was to be expected, because the conditional independence assumption explicitly underlying dichotomous IRT models (and implicit in traditional test theory) was violated. What was unexpected was the size of the effect of the unmodeled intratestlet dependence.

Traditional reading comprehension tests typically have few questions following a single passage; for example, reading comprehension items on the verbal

The order of the authors is alphabetical. This work was collaborative in every respect. It was supported by the Educational Testing Service through funding from the Program Research Planning Council and the New Possibilities Project. We would like to express our gratitude for this help. The work of Stephen Sireci was done while he was an ETS Summer Predoctoral Fellow. The SAT-V data were made available to us by Nancy Feryok, John Fremer, and Ida Lawrence. The clarity and correctness of the manuscript was helped through the comments of Bill Angoff, Bob Mislevy, and Lynne Steinberg. We would like to express our gratitude to them all. The responsibility for any errors that remain is ours alone.

part of the Scholastic Aptitude Test (SAT-V) contain passages of modest length and about four items. Among possible revisions to this test are testlets consisting of much longer passages with 9 to 14 questions per passage. Traditional reliability coefficients calculated on the older forms yield somewhat optimistic estimates; on the new forms, the overestimate is (as we will show) too large to ignore.

Historical Background

The fact that the reliability of tests built from testlets will be overestimated by item-based methods is not newly discovered. It has been well known for at least 60 years. Warnings about inflated reliability estimates are commonly carried in many introductory measurement textbooks. For example,

> One precaution to be observed in making such an odd-even split pertains to groups of items dealing with a single problem, such as questions referring to a particular mechanical diagram or to a given passage in a reading test. In this case, a whole group of items should be assigned intact to one or the other half. Were the items in such a group to be placed in different halves of the test, the similarity of the half scores would be spuriously inflated, since any single error in understanding of the problem might affect items in both halves. (Anastasi, 1961, p. 121; this warning was included in all subsequent editions as well—cf., Anastasi, 1988, p. 121)

> In some cases, several items may be unduly closely related in content. Examples of this would be a group of reading comprehension items all based on the same passage.... In that case, it will be preferable to put all items in a single group into a single half-score . . . this procedure may be expected to give a somewhat more conservative and a more appropriate estimate of reliability.... (Thorndike, 1951, p. 585)

> Interdependent items tend to reduce the reliability. Such items are passed or failed together and this has the equivalent result of reducing the length of the test. (Guilford, 1936, p. 417)

These warnings were generated by an earlier exchange about the potential usefulness of the Spearman-Brown prophecy formula (cited in Gulliksen, 1950/1987). This exchange is summarized by Brown and Thomson (1925), who reported criticisms of the mathematician W. L. Crumm (1923). Crumm felt that the requirements of Spearman-Brown were too stringent to expect them to be met in practice. Holzinger (1923) reported some empirical evidence supporting the Spearman-Brown formula, and then Truman Kelley (1924) provided a fuller mathematical defense of Spearman-Brown. Of interest here is Kelley's (1924) comment, "If two or more exercises contain common features, not found in the general field, then the Spearman-Brown r_{II} will tend on this account to be too large" (p. 195).

All of these are in the context of even-odd split-half reliability; but coefficient α averages all of the split-half reliabilities. Thus, if some of the split-half reliabilities are inflated by within-passage correlation, so too will be α.

The 1966 edition of the APA *Standards for Educational and Psychological*

Tests and Manuals (Section D5.4) deemed treating testlets in a unitary matter "essential" (p. 30). It stated:

> If several questions within a test are experimentally linked so that the reaction to one question influences the reaction to another, the entire group of questions should be treated preferably as an item when the data arising from application of split-half or appropriate analysis-of-variance methods are reported in the test manual. (p. 30)

The importance of the recommendation (Section F5.4) was diminished to "very desirable" (p. 53) in the 1974 *Standards* and, as nearly as we can tell, it was omitted entirely from the most recent (1985) *Standards.* Has the effect of local dependence on reliability estimation diminished over the decades? We think not. One purpose of this article is to reaffirm the wisdom of the earlier *Standards* (1966) by indicating the size of the error that can be made by ignoring the test structure. For a variety of reasons and purposes, we, along with our colleagues (Wainer & Kiely, 1987; Thissen et al., 1989; Wainer & Lewis, 1990; Wainer, Sireci, & Thissen, 1991) have proposed (essentially) treating testlets as items; thus, in this article, we concentrate on the effects on the estimation of reliability that arise from failing to do so.

Reliability

> *reliable* \ri-'lī-ə-bəl\ *adj* (1564) **1:** suitable or fit to be relied on: DEPENDABLE **2:** giving the same result on successive trials. . . . (*Webster's Ninth New Collegiate Dictionary,* 1987, p. 995)

Reliability is defined (Lord & Novick, 1968, p. 61) as the squared correlation ρ^2_{XT} between observed score (X) and true score (T). This can be expressed as the ratio of true score variance to observed score variance or, after a little algebra, as

$$\rho^2_{XT} = \frac{\sigma^2_X - \sigma^2_e}{\sigma^2_X}, \tag{1}$$

where the subscript *e* denotes error. In this article, we use coefficient α (see Lord & Novick, 1968, p. 204) to estimate the reliability of the summed scores of the traditional test theory.

There are many ways to calculate reliability, but in traditional test theory reliability takes a single value that describes the average error variance for all scores. (Traub & Rowley, 1991, p. 40 use the notation Ave(σ^2_e) in place of σ^2_e to emphasize that only an average estimate of the error variance is considered in true score reliability theory.) In contrast to this simplification, measurement precision in an IRT system can be characterized as a function of proficiency (θ); therefore, precision need not be represented by a single overall reliability. Precision in an IRT system is usually described in terms of $I(\theta)$, the information function, the conditional error variance σ^2_{e*} , or the standard error σ_{e*}; these also vary as functions of θ. Although measurement error may vary as a function of proficiency, Green, Bock, Humphreys, Linn, & Reckase (1984) observe that it can also be averaged to give marginal reliability comparable to that of the traditional theory. The marginal measurement error variance, $\bar{\sigma}^2_{e^*}$, for a population with proficiency density $g(\theta)$ is

$$\bar{\sigma}^2_{e^*} = \int \sigma^2_{e^*}\, g(\theta)\, d\theta, \tag{2}$$

where $\sigma^2_{e^*}$ is the expected value of the error variance associated with the expected a posteriori estimate at θ, and the marginal reliability is

$$\bar{\rho} = \frac{\sigma^2_\theta - \bar{\sigma}^2_{e^*}}{\sigma^2_\theta}\,. \tag{3}$$

Here, the integration (or averaging) over possible values of θ takes the place of the traditional characterization of an average error variance. The value of $\bar{\sigma}^2_{e^*}$ is the average of the (possibly varying) values of the expected error variance, $\sigma^2_{e^*}$; if many values of $\sigma^2_{e^*}$ were tabulated in a row for all the different values of proficiency (θ), $\bar{\sigma}^2_{e^*}$ would be that row's marginal average. Therefore, the reliability derived from that marginal error variance is called the *marginal reliability* and denoted $\bar{\rho}$ to indicate explicitly that it is an average. There is some loss of information in averaging unequal values of $\sigma^2_{e^*}$. But such marginal reliabilities for IRT scores parallel the construction of internal consistency estimates of reliability for traditional test scores.[1] In the remainder of this article, we use $\bar{\rho}$ to describe the reliability of estimates of latent proficiency (θ) based on IRT.

The surface analogy between Equation 1 and Equation 3 is slightly deceiving. The term, $\sigma^2_{e^*}$, is the expected value of the error variance of the expected a posteriori estimate of θ as a function of θ, computed from the information function for the expected a posteriori estimate of θ. As Birnbaum (1968, pp. 472ff) illustrates, any method of test scoring has an information function. Here, we use the information function for the expected a posteriori estimate. That is analogous to the error of estimation, σ^2_ϵ, in the traditional theory, not σ^2_e (see Lord & Novick, 1968, p. 67). And σ^2_θ is the true population value of the variance of θ.

The definition of traditional reliability in terms of true score variance, σ^2_T, and the error of estimation, σ^2_ϵ, may be straightforwardly derived from Lord and Novick's (1968, p. 67) Equation 3.8.4:

$$\sigma_\epsilon = \sigma_T \sqrt{1 - \rho^2_{XT}}\quad ,$$

which gives

$$\rho^2_{XT} = \frac{\sigma^2_T - \sigma^2_\epsilon}{\sigma^2_T}\quad ;$$

that is the classical analog of Equation 3.

An IRT Model for Testlets

It is possible to approach *testlet analysis* (the item analysis of testlets) using the tools of the traditional test theory, but we have found the tools of IRT to be more useful. In the computation of the IRT index of reliability, $\bar{\rho}$, we follow

Thissen et al. (1989) in their use of Bock's (1972) model. There we have J testlets, indexed by j, where $j = 1, 2, \ldots, J$. On each testlet, there are m_j questions, so that for the jth testlet there is the possibility for the polytomous response, $x_j = 0, 1, 2, \ldots, m_j$. The statistical testlet scoring model posits a single underlying (and unobserved) dimension that we call *latent proficiency* and denote θ. The model then represents the probability of obtaining any particular score as a function of proficiency. For each testlet, there is a set of functions, one for each response category. These functions are sometimes called *item response functions* (IRF; Holland, 1990).

The IRF for score $x = 0, 1, \ldots, m_j$ for testlet j is

$$T_{jx}(\theta) = \frac{\exp\,[a_{jx}\theta + c_{jx}]}{\sum_{k=0}^{m} \exp\,[a_{jk}\theta + c_{jk}]}, \tag{4}$$

where $\{a_k, c_k\}_j$, $k = 0, 1, \ldots, m_j$ are the item category parameters that characterize the shape of the individual response functions. If this model yields a satisfactory fit, information is calculated and inverted to yield estimates of the error variance function. Error variance is relative to the variance of θ, which distribution, $g(\theta)$, is fixed as $N(0,1)$ to help identify the model. The integration indicated in (2) can then be carried out, and $\bar{\rho}$ can be calculated through Equation 3.

An Example—A Candidate for a New SAT-V

The SAT-V is currently under revision. Many changes are being considered. One of these involves the incorporation of much longer reading passages coupled with many more questions. In one pretest administration, there were four passages yielding a total of 45 items (12, 13, 10, and 10 items per passage, respectively). In an investigation of differential testlet functioning for these testlets (Wainer, Sireci, & Thissen, 1991), we noted that the reliability of this reading comprehension test seemed to be much lower than it had been believed to be, based on coefficient α computed for the 45 items. Table 1 summarizes alternative computations of the reliability of this test.

Because we had previously found that a different item response model was required to fit the data for the male and female examinees for some of these reading passages, we computed reliability separately for the male ($N = 1{,}812$) and female ($N = 2{,}216$) examinees. Coefficient α was computed using SPSS-X RELIABILITY; $\bar{\rho}$ was computed using the conventional 3PL model for the test as 45 binary items, Bock's (1972) nominal model (as described above for the testlets), and the computer program *MULTILOG* (Thissen, 1991).

If the test is taken (incorrectly) to be 45 conditionally independent items, coefficient α and the IRT-based $\bar{\rho}$ range from 0.86 to 0.88. If the test is considered to comprise four conditionally independent testlets, coefficient α and the IRT-based $\bar{\rho}$ range from 0.76 to 0.80—each testlet-based reliability being 0.08–0.12 lower than the corresponding item-based estimate. (Prior to fitting the IRT model, we collapsed little-used extreme score categories on

TABLE 1
Traditional (α) and Item Response Theory ($\bar{\rho}$) indices of reliability for a reading comprehension test from the experimental SAT-V

Test as:	α	$\bar{\rho}$
45 items—males	.88	.87
45 items—females	.86	.88
4 testlets† —males	.80	.78
4 testlets† —females	.76	.76
4 testlets†† —males	.79	—
4 testlets†† —females	.75	—

†For the IRT analysis described in detail by Wainer, Sireci & Thissen (1991), extreme score categories were collapsed so that there were ten raw-score categories (0-9 items correct) per passage.

††These values of α were obtained for the uncollapsed passage scores; IRT analyses are not available for the uncollapsed passage scores.

passages one and two so that all four reading passages had 10 ordered raw scores; thus, the effective test length of the IRT-fitted testlet test was 40 items, not 45. Coefficient α and $\bar{\rho}$ for these collapsed data are shown in the middle block of Table 1. Collapsing the little-used score categories had no effect—certainly no negative effect—on reliability; the bottom panel of Table 1 shows that the coefficient α reliabilities of the uncollapsed data (scores from 0–45) are slightly lower than the corresponding reliabilities after collapsing. No IRT estimates are available for the uncollapsed data.)

Thus, failing to take into account the dependencies caused by having four sets of items, each set referring to a common passage, yields a 10–15% overestimate of reliability. This is not trivial, because an α of 0.76 is usually considered too low for an operational test. Indeed, 0.87 is probably only marginally acceptable (the reliabilities for the SAT-V usually range between 0.90 to 0.92 [Donlon & Angoff, 1971, p. 28]).

One way to assess the size of the overestimate of reliability is to use the Spearman-Brown formula (cited in Gulliksen, 1950/1987) to estimate how many testlets of this type we would need to get back the 14% we thought we had (to yield an α of 0.87, from an α of 0.76). To accomplish this, we would need to nearly double the test length (8 passages). An α of 0.92 would require tripling the length (12 passages!). Because it takes almost an hour to administer four testlets, we might infer that, if we built a test consisting solely of this kind of item, we would need three hours to obtain the level of reliability that is currently obtained in less than half that time on the SAT-V.

Another Example

Thissen et al. (1989) used the testlet approach in the analysis of a different reading comprehension test. In their example, there were four passages with 22 associated questions (7, 4, 3, and 8 items, respectively). They found that, when the testlet structure was explicitly characterized with a polytomous IRT model, it revealed that more traditional approaches overestimated the precision of measurement as characterized by the test information function. Thissen et al. (1989) did not report single-value reliability coefficients for those data; we report those values here, in Table 2.

Thissen et al. (1989) analyzed the responses of a single sample of N = 3,866 examinees; no differentiation was made by sex. Again, coefficient α was computed using SPSS-X RELIABILITY; $\bar{\rho}$ was computed using the conventional 3PL model for the test as 22 binary items, Bock's (1972) nominal model (as described for the testlets), and the computer program *MULTILOG* (Thissen, 1991). In this case, $\bar{\rho}$ is 0.04 greater than coefficient α for both test structures; this is probably due to the fact that $\bar{\rho}$ is reliability for optimally weighted scores, whereas α is the reliability of unweighted summed scores. In this example, there are substantial differences among the optimal weights for the four testlets.

However, the pattern of overestimation of reliability when the test is analyzed at the item level is identical to that observed for the experimental SAT-V form described above: The item level reliability is 0.08, or 12–13% higher, than the testlet reliability, regardless of whether one considers α or $\bar{\rho}$. Thissen et al. (1989) also provide the values shown in Table 2 as $r_{\text{concurrent}}$; those are the correlations with a concurrently administered 54-item verbal test, using IRT-based proficiency estimates from the 3PL model for the test as 22 items and Bock's (1972) nominal model for the test as four testlets. (The 54-item test comprised antonym, analogy, and sentence completion items; it was not really a

TABLE 2

Traditional (α) and Item Response Theory ($\bar{\rho}$) indices of reliability for the reading comprehension test described by Thissen, Steinberg & Mooney (1989)

Test as:	α	$\bar{\rho}$	$r_{\text{concurrent}}$†
22 items	.70	.74	.65
4 testlets	.62	.66	.66

†$r_{\text{concurrent}}$ is the correlation between the raw score on a concurrently administered 54-item verbal test and IRT-based estimates of proficiency from the 3PL model for the test as 22 items, and the nominal model for the test as 4 testlets.

parallel or alternate form. However, the correlation is so high that it is more tempting to call $r_{\text{concurrent}}$ an alternate-form reliability than it is to call it validity.) First, it is clear that the testlet-based reliability is close to (actually, the same as!) the value of the correlation of the test scores with the other test; this is not true of the overestimates based on 22 items. Second, Thissen et al. (1989) showed that the correlation with the concurrently administered 54-item test was significantly higher (albeit only 0.01) for the IRT testlet-based scores than for the IRT scores based on 22 individual items.

Conclusions

What is the right estimate of reliability? In the context of IRT, the answer to that question is straightforward. Because the item response model assumes local independence, it is the estimate of reliability computed with the model fitted to the elements of the test that produce locally independent responses. The point of (marginal) reliability is to describe the (average) posterior variance. The posterior for each item response pattern is computed by multiplying the trace lines, and that is only justified (or correct) if local independence holds between the trace lines that are multiplied. So it is very simple: If, and only if, the items are locally independent, then the product of the item trace lines is an accurate description of the posterior density for examinees with that response pattern, and item-based marginal reliability provides an accurate estimate of the (average) variance of those posterior densities. If local independence only holds between some larger units of the test (i.e., testlets), then trace lines for those units are multiplied to produce the posterior densities, and the correct estimate of (marginal) reliability is based on those trace lines.

It is difficult to make inferences about local independence in large sets of items. However, in the case of reading comprehension tests, there is factor analytic evidence of a lack of local independence (see Thissen et al., 1989), and there are also obvious sources of local dependence (e.g., prior knowledge of the topical material in the reading passage). Given these considerations, it seems clear that IRT analysis of reading comprehension tests most properly proceeds using the testlet (passage) as the unit of analysis (Thissen et al., 1989; Wainer & Kiely, 1987; Wainer & Lewis, 1990; Wainer, Sireci, & Thissen, 1991).

In the context of the traditional true score theory, local independence is not assumed, and there are no explicit posterior variances, so some other fundamental aspect of reliability must be considered to select the right estimate. Internal consistency estimates of reliability, like coefficient α, were developed as a means of estimating alternate form reliability without repeated testing. So the right internal consistency estimate of reliability is the one that would correspond to the (hypothetical) correlation between (equally hypothetical) alternate forms. Internal consistency estimates of reliability work to the extent that they accurately estimate the correlation between two forms of the test. The correlation between two test forms depends on the correlations between the items on one form and the items on the other form. Internal consistency estimates of reliability are based on a simple extrapolation from the (average) correlation among the items on one form (that one has) to the (average) correlation between those items and some other forms' items (that one does not have): The extrapolation is that the (average) correlation among the items one has is the same as the correlation of those items with the (hypothetical) second forms' items.

Given this, traditional internal consistency estimates of reliability are likely to be right if the correlations among the items on the test one has are representative of the correlations one would find if one produced a second form and correlated the items on that form with those on the test one has. Many item-based tests work like that. However, many tests (comprising what an item response theorist would call testlets), such as reading comprehension tests, do not work that way. In a reading comprehension test, the item responses within-passage are more highly correlated than are item responses between testlets. Therefore, if internal consistency reliability is computed at the item level, the two levels of correlation (within-passage and between-passage) are averaged—and the result is generally a higher value than the average correlation that would be obtained between items on one reading-comprehension test and items on an alternate form, because on the alternate form all of the correlations would be of the (lower) between-passage variety. That is the effect that has been illustrated in this article.

Therefore, for the traditional true-score theorist, the right estimate of reliability is computed using as units of analysis the components of the test whose average correlation represents the average correlations those components would have with their corresponding elements on an alternate form of the test. This is the point Kelley (1924), Guilford (1936), Thorndike (1951), and Anastasi (1961), quoted in the introduction to this article, were trying to make.

We fully expect that in the future tests will be based upon larger tasks—what we have called testlets. One goal of such tests is to mirror more closely what are conceived to be real world tasks. It is important that the statistical models used to analyze the data obtained from such tests be flexible enough to match the test's structure. We have shown that it is important to appropriately estimate test reliability in such a circumstance. In our examples, we have shown how treating such tests as sets of independent items yields highly misleading results—specifically, the reliability appears to be as high as what would in fact have been the case had we doubled the test's length. Such an overestimate of reliability is too large to ignore.

Note

[1] Technically, R. Mislevy (personal communication, July 11, 1990) has pointed out that if the information function is steeper than the proficiency distribution the integral in (2) can become unbounded. This is typically not a problem in applications where one approximates the integral with a sum. Yet one could imagine a case in which changing the bounds of the integral could drastically alter results. Mislevy's alternative is to

integrate the information function directly and then invert. Specifically, substitute for (2) the expression

$$\bar{I}(\theta) = \int_{-\infty}^{\infty} I(\theta)\, g(\theta)\, d\theta = \frac{1}{\overline{\sigma}^2_{e^*}}. \qquad (2^*)$$

Mislevy's formulation has obvious theoretical advantages, but we have too little experience as yet to indicate under what circumstances the extra protection is needed.

References

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: Author.

American Psychological Association. (1974). *Standards for educational and psychological tests.* Washington, DC: Author.

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Anastasi, A. (1961). *Psychological testing* (2nd ed.). New York: Macmillan.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392–479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37,* 29–51.

Brown, W., & Thomson, G. H. (1925). *The essentials of mental measurement* (3rd ed.). London: Cambridge University Press.

Crumm, W. L. (1923). Note on the reliability of a test, with a special reference to the examinations set by the College Entrance Board. *The American Mathematical Monthly, 30*(6), Sept.-Oct. 1923.

Donlon, T. F., & Angoff, W. H. (1971). *The Scholastic Aptitude Test.* In W. H. Angoff (Ed.), *The College Board Admissions Testing Program* (pp. 15–47). New York: College Entrance Examination Board.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347–360.

Guilford, J. P. (1936). *Psychometric methods* (1st ed.). New York: McGraw-Hill.

Gulliksen, H. O. (1987). *Theory of mental tests.* Hillsdale, NJ: Lawrence Erlbaum Associates. (Original work published in 1950)

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–601.

Holzinger, K. (1923). Note on the use of Spearman's prophecy formula for reliability. *The Journal of Educational Psychology, 14,* 302–305.

Kelley, T. L. (1924). Note on the reliability of a test: A reply to Dr. Crumm's criticism. *The Journal of Educational Psychology, 15,* 193–204.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mish, F. C. (Editor in Chief). (1987). *Webster's ninth new collegiate dictionary.* Springfield, MA: Merriam-Webster.

Thissen, D. (1991). *MULTILOG user's guide (Version 6)* [Computer program]. Mooresville, IN: Scientific Software.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26,* 247–260.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice, 10,* 37–45.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1–14.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement,* 28, 197–219.

Authors

STEPHEN G. SIRECI is Psychometrician, American Institute of Certified Public Accountants, Examinations Division, 1211 Avenue of the Americas, New York, NY 10036. *Degrees:* BA, MA, Loyola College; PhD Candidate, Fordham University. *Specializations:* psychometrics and classification.

DAVID THISSEN is Professor of Psychology, University of North Carolina, CB 3270, Davie Hall, Chapel Hill, NC 27599-3270. *Degrees:* BA, St. Louis University; PhD, University of Chicago. *Specializations:* quantitative psychology and measurement.

HOWARD WAINER is Principal Research Scientist, Educational Testing Service, and Professor (visiting) of Civil Engineering and Psychology, Princeton University, Princeton, NJ 08541. *Degree:* PhD, Princeton University. *Specializations:* statistical graphics, psychometrics, and computerized testing.