Researchers too frequently consider the reliability of the scores they analyze, and this may lead to incorrect conclusions. Practice in this regard may be negatively influenced by telegraphic habits of speech implying that tests possess reliability and other measurement characteristics. Styles of speaking in journal articles, in textbooks, and in professional standards and guidelines are explored. Two recommendations are offered. First, the statement "the test is reliable" should be recognized as being inappropriate, and professional standards and editorial guidelines should make this clear. Second, an important implication of the realization that reliability inures to data, rather than tests, is that reliability should generally be explored whenever data are collected. Three tables present language usage examples. An appendix lists 52 articles surveyed. (Contains 28 references.) (Author/SLD)
It is Incorrect to Say "The Test is Reliable":
Bad Language Habits
Can Contribute to Incorrect or Meaningless Research Conclusions

Bruce Thompson
Texas A&M University 77843-4225
and
Baylor College of Medicine

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX, January 27, 1994. The author acknowledges the thoughtful comments of Jerome T. Kapes on a previous version of this paper.
ABSTRACT

Researchers too infrequently consider the reliability of the scores they analyze, and this may lead to incorrect conclusions. Practice in this regard may be negatively influenced by telegraphic habits of speech implying that tests possess reliability and other measurement characteristics. Styles of speaking in journal articles, in textbooks, and in professional standards and guidelines, are explored. Suggestions for improved practice are presented.
Most of us, in both our daily lives and in our scholarship, are guided in our behavior by our paradigms. As defined by Gage (1963, p. 95), "Paradigms are models, patterns, or schemata. Paradigms are not the theories; they are rather ways of thinking or patterns for research." Tuthill and Ashton (1983, p. 7) explained that:

A scientific paradigm can be thought of as a socially shared cognitive schema. Just as our cognitive schema provide us, as individuals, with a way of making sense of the world around us, a scientific paradigm provides a group of scientists with a way of collectively making sense of their scientific world.

But scholars usually do not consciously recognize the influence of their paradigms. As Lincoln and Guba (1985, pp. 19-20) noted:

If it is difficult for a fish to understand water because it has spent all its life in it, so it is difficult for scientists... to understand what their basic axioms or assumptions might be and what impact those axioms and assumptions have upon everyday thinking and lifestyle.

Even though social scientists are usually unaware of paradigm influences, paradigms nevertheless are potent influences in that they tell us what we need to think about, and also the things about which we need not think. As Patton (1975, p. 9) suggested,
Paradigms are normative, they tell the practitioner what to do without the necessity of long existential or epistemological consideration. But it is this aspect of a paradigm that constitutes both its strength and its weaknesses--its strength in that it makes action possible; its weakness in that the very reason for action is hidden in the unquestioned assumptions of the paradigm.

Although scholars are usually blind to the impacts of their paradigms, occasionally paradigm presumptions "leak out" in the language that scientists use. Conversely, the things we say conventionally, even when our jargon has become telegraphic shorthand, can subsequently come to be perceived by us as literal truth, and then unquestioned, within the context of our paradigms.

One common feature of contemporary scholarly language is the usage of the statement, "the test is reliable." The purpose of this essay is to argue that such language is both incorrect and deleterious in its affects on scholarly inquiry, particularly given the pernicious consequences that unconscious paradigmatic beliefs can exact.

The paper the nature of reliability is reviewed, and then the consequences of insufficiently considering reliability when conducting substantive research addressing basic and applied problems is considered. Next, language use in one prominent journal is reviewed, related language use in four prominent textbooks is reviewed, and then language use in profesional
standards and guidelines is considered. Finally, suggestions for improved practice are presented.

The Nature of Reliability

Too few researchers act on a conscious recognition that reliability is a characteristic of scores or the data in hand. Many authors present this view, but paradigm influences constrain some researchers from actively integrating this presumption into their actual analytic practice.

As Rowley (1976, p. 53, emphasis added) argued, "It needs to be established that an instrument itself is neither reliable nor unreliable.... A single instrument can produce scores which are reliable, and other scores which are unreliable." Similarly, Crocker and Algina (1986, p. 144, emphasis added) argued that, "...A test is not 'reliable' or 'unreliable.' Rather, reliability is a property of the scores on a test for a particular group of examinees."

In another widely respected text, Gronlund and Linn (1990, p. 78, emphasis in original) noted,

Reliability refers to the results obtained with an evaluation instrument and not to the instrument itself.... Thus, it is more appropriate to speak of the reliability of the "test scores" or of the "measurement" than of the "test" or the "instrument."

And Eason (1991, p. 84, emphasis added) argued that:

Though some practitioners of the classical
measurement paradigm [incorrectly] speak of reliability as a characteristic of tests, in fact reliability is a characteristic of data, albeit data generated on a given measure administered with a given protocol to given subjects on given occasions.

The subjects themselves impact the reliability of scores, and thus it becomes an oxymoron to speak of "the reliability of the test" without considering to whom the test was administered, or other facets of the measurement protocol. Reliability is driven by variance--typically, greater scores variance leads to greater score reliability, and so more heterogeneous samples often lead to more variable scores, and thus to higher reliability. Therefore, the same measure, when administered to more heterogenous or to more homogeneous sets of subjects, will yield scores with differing reliability. As Dawes (1987, p. 486) observed, "...Because reliability is a function of sample as well as of instrument, it should be evaluated on a sample from the intended target population--an obvious but sometimes overlooked point."

Our shorthand ways of speaking (e.g., language saying "the test is reliable") can itself cause confusion and lead to bad practice. As Pedhazur and Schmelkin (1991, p. 82, emphasis in original) observed, "Statements about the reliability of a measure are... inappropriate and potentially misleading." These telegraphic ways of speaking are not inherently problematic, but they often later become so when we come unconsciously to ascribe literal truth to our shorthand, rather than recognizing that our
jargon is sometimes telegraphic and is not literally true. As noted elsewhere:

This is not just an issue of sloppy speaking--the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice.

Thompson (1992, p. 436)

The Important Impacts of Reliability on Substantive Research

In one book exploring the intimate linkages between measurement error variance and our attributions about the origins of variance in our substantive basic or applied research research, Pedhazur and Schmelkin (1991) noted,

Measurement error is the Achilles' heel of sociobehavioral research. Although most programs in sociobehavioral sciences, especially doctoral programs, require a modicum of exposure to statistics and research design, few seem to require the same where measurement is concerned. Thus, many students get the impression that no special competencies are necessary for the development and use of measures... (pp. 2-3)

Therefore, it should not be surprising that studies of research reports in journals indicate insufficient attention to the impacts of measurement integrity on the integrity of substantive research conclusions. For example, with respect to the American
Educational Research Journal, Willson (1980) reported that:

...Only 37% of the AERJ studies explicitly reported reliability coefficients for the data analyzed. Another 18% reported only indirectly through reference to earlier research.... That reliability... is unreported in almost half the published research is... inexcusable at this late date...." (pp. 8-9)

A more recent "perusal of contemporary psychology journals demonstrates that quantitative reports of scale reliability and validity estimates are often missing or incomplete" (Meier & Davis, 1990, p. 113); and that "the majority [95%, 85% and 60%] of the scales described in the [three Journal of Counseling Psychology] JCP volumes [1967, 1977 and 1987] were not accompanied by reports of psychometric properties" (p. 115). The situation is apparently roughly equivalent as regards dissertation research (Thompson, 1988).

This state of affairs is surprising, given two related trends within the literature. First, since the influential articles by Cohen (1968) and Knapp (1978) appeared, more researchers have recognized that all parametric statistical analyses are correlational (Thompson, 1991), and that substantive variance-accounted-for effect sizes expressed as $r^2$ analogs can be interpreted in all studies. Second, the importance of interpreting effect sizes as against statistical significance tests has been increasingly recognized (e.g., Thompson, 1993), as reflected, for

Nevertheless, too few researchers act on the premise that score reliability establishes a ceiling for substantive effect sizes. These impacts can be readily illustrated in a concrete example using the bivariate correlation as an heuristic.

It has been recognized in textbooks dating back to the 1950s, and in more recent books as well (e.g., Pedhazur & Schmelkin, 1991, p. 114), that a correlation coefficient "corrected" for attenuation due to measurement error ($\hat{r}_{XY}$) can be estimated as:

$$\hat{r}_{XY} = \frac{r_{XY}}{\left[ (r_{XX} \times r_{YY}) \right]^{1/2}}$$

where $r_{XY}$ is the calculated bivariate relationship between scores on variables $X$ and $Y$, and $r_{XX}$ and $r_{YY}$ are respectively the reliability coefficients for scores on $X$ and $Y$. This algorithym can be re-expressed in the more familiar metric of common variance, as is often done in popular variance-accounted-for effect size statistics (e.g., $\hat{R}^2$, $R^2$, eta$^2$, omega$^2$):

$$\hat{r}_{XY}^2 = \frac{r_{XY}^2}{(r_{XX} \times r_{YY})}$$

Through algebraic manipulation, the detectable effect size, given knowledge of "true" relationship, $\hat{r}_{XY}^2$, and the reliabilities of the two sets of scores, is:

$$r_{XY}^2 = \hat{r}_{XY}^2 \times (r_{XX} \times r_{YY})$$

Even if the "true" relationship between perfectly reliable measures of $X$ and $Y$ was perfect, i.e., $\hat{r}_{XY}^2 = 1.0$, the detectable effect in any study can never exceed the product of the reliability
coefficients for the two sets of scores:

\[ r_{XY}^2 = 1 \times (r_{XX} \times r_{YY}) \]

For example, even when \( r_{XY}^2 = 1.0 \), if both sets of scores have reliability coefficients of .7, the detectable effect cannot exceed .49. Clearly, measurement error prospectively impacts the effect size that we can obtain in a planned study and also should be retrospectively considered when interpreting calculated effects once the study has been done.

The failure to consider score reliability in substantive research may exact a toll on the interpretations within research studies. We may conduct studies that could not possibly yield noteworthy effect sizes. Or we may not accurately interpret our results if we do not consider the reliability of the scores we are actually analyzing.

These practices may be caused by misperceptions that tests can be reliable or valid. These misperceptions themselves may be caused, or at least reinforced, by the use of telegraphic language that comes to be unconsciously believed as literal truth, and then unconsciously incorporated into paradigms for behavior.

Language Use in A Prominent Measurement Journal

Logically, if the language used by the best experts to describe measurement integrity was telegraphic or inappropriate, then, a fortiorari, appropriate language use and thinking by others regarding score reliability would be even less likely. One empirical snapshot of contemporary language practice was derived for the present paper by reviewing all the articles in the
measurement integrity studies section of Educational and Psychological Measurement (EPM). EPM is a journal that was started some 50 years ago with Frederick Kuder as Founding Editor, and until very recently, also as the journal's owner. Kuder, of course, is widely known for his contributions to reliability theory through the various "KR" formulas.

The 1992 volume of EPM contained 64 articles in the journal's measurement integrity section. Eleven of these articles did not directly deal with measurement characteristics issues. One of the remaining 53 articles involved the present author as a coauthor, and did not involve the language use issues described here. Table 1 presents illustrative quotations from the remaining 52 articles. The tabled quotations, even in a respected forum presumably involving measurement experts as authors and reviewers, reflect a pattern of language usage regarding measurement characteristics that is at best telegraphic in style.

INSERT TABLE 1 ABOUT HERE.

Language Use in Four Prominent Measurement Texts

Four well-known measurement textbooks (Gronlund & Linn, 1990; Mehrens & Lehmann, 1991; Sax, 1989; Thorndike, Cunningham, Thorndike, & Hagen, 1991) were also surveyed to garner an impression of language use as regards score reliability. Table 2 presents illustrative quotations from these works. Even respected texts being published in as late as 6th editions reflect language usage that is at best inconsistent, telegraphic, or incorrect.
One set of authors, for example, presents an oxymoron in which it is asserted that (a) the sample impacts reliability but that (b) somehow over different samples still "the test is reliable". These authors note, "A third factor influencing the estimated reliability of a test is group homogeneity" (Mehrens & Lehmann, 1991, p. 259, emphasis added).

Language Use in Professional Standards and Guidelines

The language in professional journals and textbooks has both influenced and been influenced by the language use in professional standards and guidelines. For example, Meier and Davis (1990, p. 113) suggested that so few authors may test or even discuss the reliability of their scores partially as

...the result of a lack of explicit guidelines for the reporting of scale information. For example, the Publication Manual of the American Psychological Association (American Psychological Association, 1983) makes no specific recommendations in regard to the reporting of scales' psychometric properties.

Table 3 reports related language use in two fairly recent sets of professional standards (APA/AERA/NCME, 1985; Joint Committee, in press). For example, the APA/AERA/NCME (1985) test standards emphasize that, "Because there are many ways of estimating reliability, each influenced by different sources of measurement error, it is unacceptable to say simply, 'The reliability of test
X is .90"" (p. 21). Yet, on the same page, these standards speak of "the reliability of a highly speeded test" (APA/AERA/NCME, 1985, p. 21, emphasis added).

Conclusions

Based on these considerations, two recommendations are offered. First, the language of saying "the test is reliable" should be recognized as being inappropriate, and professional standards and editorial guidelines should make forcefully this clear. Instead, authors should be encouraged to say, "the scores in our study had a classical theory test-retest reliability coefficient of X," or "based on generalizability theory analysis, the scores in our study had a phi coefficient of X."

It will not be sufficient to say in our standards that, "Because there are many ways of estimating reliability, each influenced by different sources of measurement error, it is unacceptable to say simply, 'The reliability of test X is .90'" (APA/AERA/NCME, 1985, p. 21). Rather, such language usage should be declared inappropriate because the language is, on its face, untrue. And the consequences of believing untrue shorthands should be noted within our professional standards.

Of course, the illustrations of language use presented in Tables 1 through 3 suggest that changing our habits of speech will be a daunting task. But, as Lachman (1993) noted, "Language habits are difficult to change. Sometimes, however, it is appropriate and
desirable to change them" (p. 1093).

Second, as suggested elsewhere,

One important implication of the realization that reliability inures to data (rather than tests) is that reliability should generally be explored whenever data are collected. And we always need to thoughtfully and explicitly explore whether the data in hand were collected on a sample similar to the samples used in previous reliability studies with a given measure. (Thompson, 1992, p. 436)

Such practices would provide better models for behavior, would provide more information in the literature about the data from our measures, and would themselves challenge paradigmatic assumptions that "the test is [or can be] reliable."
References
Joint Committee on Standards for Educational Evaluation. (in press). The program evaluation standards: How to assess
evaluations of educational programs. Newbury Park, CA SAGE.


American Psychologist, 44, 1276-1284.
Table 1
Illustrative Journal Quotations Illustrating Telegraphic Language (Emphases Added to All Quotations)

**Examples of Telegraphic Misspeaking**

"The Speed of Thinking Test appears to provide a measure of cognitive speed that is sufficiently reliable and valid..." (Carver, 1992, p. 132)

"...a major shortcoming of research in this domain has been the lack of a reliable and valid measure..." (Schriesheim, Neider, Scandura & Tepper, 1992, p. 136)

"The internal consistency reliability for the scales and subscales were calculated using Cronbach's alpha." (Caruso, 1992, p. 156)


"...the internal consistency reliabilities (coefficient alpha) of the new scales were computed..." (Romero, Tepper & Tetrault, 1992, p. 176)

"The results of this study suggest that the scale developed here is highly reliable..." (Murphy & Thorton, 1992, p. 199).

"...the SWMSS possessed strong reliability, and convergent and discriminant validity..." (Vandenberg & Scarpello, 1992, p. 204)

"...Cronbach's alpha showed that the overall reliability of the 20-item scale was..." (Chow & Winzer, 1992, p. 227)

"Evidence on the reliability, stability, and validity of the NEO-PI has been reviewed..." (McCrae & Costa, 1992, p. 232)

"The results of the statistical analyses indicate that the Student Religiosity Questionnaire provides a reliable measure..." (Katz & Schmida, 1992, p. 355)

"The concurrent validity of the MTA scale was supported..." (d'Ailly & Bergering, 1992, p. 370)

"...a lack of predictive validity of this subtest in medical education." (Glaser, Hojat, Veloski, Blacklow & Goepp, 1992, p. 405)

"Reliability of the 20-item scale was determined using coefficient alpha..." (Smither & Houston, 1992, p. 414)

"The instrument used to measure comprehension monitoring ability was found to have substantial reliability..." (Otero, Campanario & Hopkins, 1992, p. 428)
"Reliability data... show adequate to high alpha coefficients to each subscale..." (Thornburg, Ispa, Adams & Lee, 1992, p. 432)

"...most of these developing abilities were also the ones that had high 4-year test-retest reliabilities..." (Dawis, Goldman & Sung, 1992, p. 464)

"One approach to determining construct validity of a test is to examine item content..." (Wooley & Hakstian, 1992, p. 476)

"...the items are more valid for men than for women." (Novy, 1992, p. 494)

"...this measure is not reliable..." (Rentsch & Heffner, 1992, p. 646)

"...a reliable... measure of computer attitudes among professional nurses." (Coover & Delcourt, 1992, p. 654)

"...establish the construct validity of a psychometric instrument for assessing beliefs..." (Silvernail, 1992, p. 667)

"The validity of such instruments..." (Austin, 1992, p. 669)

"After examining four inventories, Biaggio (1980) questioned their construct validity..., their poor reliability, and limited predictive validity." (Kroner, Reddon & Serin, 1992, p. 688)

"...the shorter scales are a little less reliable than the longer scales..." (Francis & Katz, 1992, p. 697)

"...the comparative validity of the two measures..." (Goldstein & Bokoros, 1992, p. 707)

"...reliabilities of the item sets were moderate..." (Beyler & Schmeck, 1992, p. 713)

"If the subtests weighted in this process were not valid..." (Earles & Ree, 1992, pp. 722)

"...the reliability and validity... of two American-developed instruments..." (Watkins & Gerong, 1992, p. 728)

"The obtained estimates of internal-consistency reliability for the Revised Maslach Burnout Scale was .82..." (Gryskiewicz & Buttner, 1992, p. 749)

"The internal-consistency reliability coefficient (coefficient alpha) for the scale was 0.90.... It would also appear to be a valid instrument..." (Pretorius & Norman, 1992, pp. 936-937)
"...the construct validity of the original LSI..." Geiger, Boyle, & Pinto, 1992, p. 758)

"...it is important that a valid measure be found..." (Gold, Roth, Wright, Michael & Chen, 1992, p. 762)

"...whether or not the predictive validity of the Leniency Scale would be affected..." (Highhouse, 1992, p. 785)

"These achievement tests have reliability estimates greater than .92." (Marjoribanks, 1992, p. 947)

"...the lack of valid and reliable instruments..." (Short & Rinehart, 1992, p. 953)

"Once the reliability of the Anxiety Scale had been established..." (Sánchez-Herrero & Sánchez, 1992, p. 964)

"...the Cultural Literacy Test is very reliable..." (Pentony, 1992, p. 970)

"...question the validity of the instrument..." (Ayers & Quattlebaum, 1992, p. 973)

"Both of these scales... have evidence supporting their reliability and validity..." (Schriesheim, Scandura, Eisenbach & Neider, 1992, p. 985)

"With respect to the reliability of the scale, results from this study revealed that the internal consistency of all subscales was adequate..." (Vallerand, Pelletier, Blais, Brière, Senécal & Vallières, 1992, p. 1015)

"...the test has demonstrated high reliability..." (Goldberg & Alliger, 1992, p. 1022)

"The two halves of the SCT have internal-consistency estimates of reliabilities greater than .80. (Novy & Francis, 1992, p. 1038)

"Cronbach's alpha for the SL-ASIA was found to be .91..." (Suinn, Ahuna & Khoo, 1992, p. 1043)

"...the SAT has even less incremental validity than their results suggest..." (Baron & Norman, 1992, p. 1054)

Anthropometric Attribution to Tests Being Actors
"The three satisfaction instruments in the study displayed reasonable levels of internal consistency reliability." (Rentsch & Steel, 1992, p. 360)

"...this shortened evaluation instrument demonstrates very high reliability..." (Fernández & Mateo, 1992, p. 679)
"The obtained factor solutions and resulting reliability coefficients for the CAS, CARS, and CSE suggest that each instrument exhibits construct validity and reliability." (Harrison & Rainer, 1992, p. 744)

**Measurement Characteristics Ascribed to Model/Theory**

"Further studies are needed to shed light on the validity of the Crites model..." (Westbrook & Sanford, 1992, p. 351)

**Inconsistent Use of Language**

"The reliability coefficients for the creativity composites [i.e., scores] were... The reliability coefficients for the Intelligence ratings were..." (Runco & Mraz, 1992, p. 217)

*versus*

"The new scoring technique... has demonstrated reliability." (Runco & Mraz, 1992, p. 219)

"Internal-consistency estimates of reliability for the total score across the grade levels is adequate..." (Hagborg & Wachman, 1992, p. 438)

*versus*

"...the validity of the instrument was supported..." (Hagborg & Wachman, 1992, p. 438)

"The reliability and validity of obtained raw scores were virtually unaffected..." (Simpson & Halpin, 1992, p. 468)

*versus*

"...no accompanying loss in reliability or validity of the test..." (Simpson & Halpin, 1992, p. 468)

"The K-BIT manual reports an internal consistency coefficient of .92 for the total sample and test-retest reliability coefficients greater than .90 for each age group." (Prewitt, 1992, p. 979)

*versus*

"...the K-BIT should have evidence supporting its concurrent validity..." (Prewitt, 1992, p. 977)

**Note.** The reference list of these and other EPM articles surveyed is available from the author upon request.
Table 2
Illustrative Book Quotations Illustrating Language Use

(Thorndike, Cunningham, Thorndike, & Hagen, 1991)
"...The larger a sample of a person's behavior we have, the more reliable the measure will be." (Thorndike, Cunningham, Thorndike, & Hagen, 1991, p. 100, emphasis added)
"...the test with the higher reliability coefficient..." (Thorndike, Cunningham, Thorndike, & Hagen, 1991, p. 100, emphasis added)
"...we prefer the more reliable test." (Thorndike, Cunningham, Thorndike, & Hagen, 1991, p. 105, emphasis added)
"...to evaluate the reliability of a test..." (Thorndike, Cunningham, Thorndike, & Hagen, 1991, p. 118, emphasis added)
"...the correct reliability for any instrument." (Thorndike, Cunningham, Thorndike, & Hagen, 1991, p. 120, emphasis added)
"How reliable a test must be..." (Thorndike, Cunningham, Thorndike, & Hagen, 1991, p. 120, emphasis added)

(Gronlund & Linn, 1990)
"Any particular instrument may have a number of different reliabilities..." (Gronlund & Linn, 1990, p. 78, emphasis added)
"...constructing more reliable classroom tests." (Gronlund & Linn, 1990, p. 93, emphasis added)
"...the reliability of their own classroom tests." (Gronlund & Linn, 1990, p. 93, emphasis added)
"In general, the longer the test is, the higher its reliability will be." (Gronlund & Linn, 1990, p. 93, emphasis added)
"...effect on the reliability of the measures obtained..." (Gronlund & Linn, 1990, p. 97, emphasis added)
"...classroom tests of questionable reliability..." (Gronlund & Linn, 1990, p. 100, emphasis added)
versus
"...for estimating the reliability of test scores." (Gronlund & Linn, 1990, p. 83, emphasis added)
"...in estimating the reliability of test scores..." (Gronlund & Linn, 1990, p. 86, emphasis added)
"...provide more reliable results..." (Gronlund & Linn, 1990, p. 93, emphasis added)
"...the reliability of the test results..." (Gronlund & Linn, 1990, p. 97, emphasis added)
"...the reliability of our criterion-referenced interpretations with these tests." (Gronlund & Linn, 1990, p. 100, emphasis added)
"In interpreting and using reliability information, it is important to remember that reliability estimates refer to the results of measurement..." (Gronlund & Linn, 1990, p. 103, emphasis in original)

(Mehrens & Lehmann, 1991)
"...No measure is perfectly reliable." (Mehrens & Lehmann, 1991, p. 249, emphasis added)
"...should result in a reasonably reliable test." (Mehrens &
Lehmann, 1991, p. 249, emphasis added)
"...estimate the reliability of their own instruments..." (Mehrens & Lehmann, 1991, p. 249, emphasis added)
"In physical measurement we can ordinarily obtain very reliable measures." (Mehrens & Lehmann, 1991, p. 249, emphasis added)
"...an estimate of the reliability (or interindividual variability) of the measure." (Mehrens & Lehmann, 1991, p. 250, emphasis in original)
"...the more consistent (reliable) the measurement." (Mehrens & Lehmann, 1991, p. 250, emphasis added)
"...estimates of the reliability of their classroom tests." (Mehrens & Lehmann, 1991, p. 256, emphasis added)
"...to estimate what the reliability of a test would be..." (Mehrens & Lehmann, 1991, p. 258, emphasis added)
"...if a test has an original reliability..." (Mehrens & Lehmann, 1991, p. 258, emphasis added)
"Just as adding equivalent items makes a test score more reliable, so deleting equivalent items makes a test less reliable." (Mehrens & Lehmann, 1991, p. 258, emphasis added)
"...a test with low reliability..." (Mehrens & Lehmann, 1991, p. 263, emphasis added)
"...complained about standardized tests because they lack perfect reliability." (Mehrens & Lehmann, 1991, p. 264, emphasis added)

"Technically speaking, data should be reliable, and the inferences we draw from the data should be valid." (Mehrens & Lehmann, 1991, p. 248, emphasis added)
"...the reliability of a set of scores." (Mehrens & Lehmann, 1991, p. 248, emphasis added)
"...the reliability of the sum (or average) of the two readers' scores." (Mehrens & Lehmann, 1991, p. 257, emphasis added)
"...longer tests give more reliable scores." (Mehrens & Lehmann, 1991, p. 258, emphasis added)
"The reliability of the data..." (Mehrens & Lehmann, 1991, p. 262, emphasis added)
"...the data should be fairly reliable..." (Mehrens & Lehmann, 1991, p. 262, emphasis added)
"...the reliability of the test..." (Mehrens & Lehmann, 1991, p. 262, emphasis added)
"...the reliability of the scores is of more concern..." (Mehrens & Lehmann, 1991, p. 263, emphasis added)
"...the scores should be more reliable..." (Mehrens & Lehmann, 1991, p. 263, emphasis added)
"...consider the quality of the data. Reliability is one of the more important qualities." (Mehrens & Lehmann, 1991, p. 264, emphasis added)

(Sax, 1989)
"Unreliable tests measure the effects of chance..." (Sax, 1989, p. 259, emphasis added)
"A test with low reliability..." (Sax, 1989, p. 259, emphasis added)
"...consideration of the reliability of measurements. Unreliable tests are no better..." (Sax, 1989, p. 259, emphasis added) 'reliability of a test' should always be interpreted to mean the 'reliability of measurements or observations derived from a test.'" (Sax, 1989, pp. 263-264, emphasis in original) "Parallel [test] forms are never perfectly correlated or reliable." (Sax, 1989, p. 264, emphasis added)

versus

"...It is more accurate to talk about the reliability of measurements (data, scores, and observations) than the reliability of tests (questions, items, and other tasks). Any reference to the "...the reliability of measurements..." (Sax, 1989, p. 273, emphasis added)

"...total scores usually have higher reliabilities." (Sax, 1989, p. 275, emphasis added)
Table 3
Language Usage in Professional Standards

(Joint Committee on Standard for Educational Evaluation, in press)

[A common error is] "[f]ailing to take into account the fact that the reliability of the scores provided by an instrument or procedure may fluctuate depending on how, when, and to whom the instrument or procedure is administered." (Joint Committee on Standard for Educational Evaluation, in press, emphasis added)

"A generic term, reliability refers to the degree of consistency of the information obtained from an information gathering process." (Joint Committee on Standard for Educational Evaluation, in press)

"Whenever possible, evaluators should choose information gathering procedures that have, in the past, yielded data and information with acceptable reliability for their intended uses; however, the generalizability of previous favorable reliability results may not be simply assumed. Reliability information should be collected that is directly relevant to the groups and ways in which the information gathering procedures will be used in the evaluation." (Joint Committee on Standard for Educational Evaluation, in press)

(APA/AERA/NCME, 1985)

"Reliability refers to the degree to which test scores are free from errors of measurement." (APA/AERA/NCME, 1985, p. 19)

"Measurement errors reduce the reliability (and therefore the generalizability) of the score obtained for a person..." (APA/AERA/NCME, 1985, p. 19, emphasis added)

"But scores representing differences between scores obtained from two tests or from repeated administrations of the same test... are generally less reliable than either of the parts." (APA/AERA/NCME, 1985, p. 20, emphasis added)
APPENDIX A

List of Volume 52 EPM Articles (n=53+11=64) Surveyed

Studies of (n=53) Measurement Characteristics


**Studies (n=11) NOT Investigating Measurement Characteristics**


