THE RELIABILITY OF MEASURING INSTRUMENTS

Thomas R. Knapp

PREFACE

Can you say "reliability" without saying "validity"? (Can you say "Rosencrantz" without saying "Guildenstern"?) I hope so, because this book is all about reliability, except for five appendices in which I discuss validity and for occasional comments in the text proper regarding the difference between reliability and validity. But isn't validity more important than reliability? Of course; a reliable instrument that doesn't measure what you want it to measure is essentially worthless. The problem is that the validity of a measurement device ultimately relies on the subjective judgment of experts in the field (all of the current emphasis on construct validity to the contrary notwithstanding), and my primary purpose in writing this book is to pursue those statistical features of measuring instruments that tell you whether or not, or to what extent, such instruments are consistent.

There are 14 chapters in the book. Chapter 1 is an introductory treatment of the concept of reliability, with special attention given to its many synonyms and nuances. The following chapter addresses the associated concept of measurement error, with an extended discussion of "randomness". Chapter 3 is devoted to classical reliability theory and is the most technical section of the book, but if you think back to your high school mathematics you will recognize the similarity to plane geometry, with its counterpart definitions, axioms, and theorems. (It is assumed that you are also familiar with descriptive statistics such as means, variances, and correlation coefficients, and with the basic principles of inferential statistics.)

Chapters 4 and 5 treat, respectively, the concept of attenuation and the interpretation of individual measurements. In Chapter 6 I try to summarize the literature regarding the reliability of difference scores of various types and the controversies concerning some of those types.

The matter of the reliability of individual test items is explored in Chapter 7. Discussion of the internal consistency reliability of the total score on a test that consists of more than one item (the usual case) follows naturally in Chapter 8, where the primary emphasis is on coefficient alpha (Cronbach's alpha). That chapter (Chapter 8) also includes a brief section in which I point out the methodological equivalence of internal consistency reliability and both inter-rater and intra-rater reliability.

Chapter 9 on intraclass correlations is my favorite chapter. Although their principal application has been to the reliability of ratings, they come up in all sorts of interesting contexts, including those concerned with the unit of analysis and the independence of observations.

Relative agreement vs. absolute agreement and ordinal vs. interval measurement provide the focus of Chapter 10. Most discussions of instrument reliability are concerned with the relative agreement between two equal-status operationalizations of a particular construct, but some are devoted exclusively to absolute agreement. Likert-type scales and other instruments that do not have equal units require special considerations. (Some of this material was originally included in various other chapters in previous editions of this book.)

Chapter 11 is concerned mostly with statistical inferences from samples of "measurees" to populations of "measurees", but some attention is also given to statistical inferences from samples of “measurers” to populations of “measurers”.

In Chapter 12 I try to bring everything together by applying classical reliability theory to a set of data that were generated in a study of alternative ways of measuring height. (The data, which have been graciously provided to me by Dr. Jean K. Brown, Dean, School of Nursing, University at Buffalo, State University of New York, are in Appendix A.)

The following chapter (Chapter 13) deals with a variety of special topics regarding instrument reliability. And a final chapter (Chapter 14) attempts to extend the concept of reliability of measuring instruments to the reliability of claims.

There is an appendix (Appendix B) on the validity of measuring instruments in general, an appendix (Appendix C) on the reliability and validity of birth certificates and death certificates, an appendix (Appendix D) on the reliability and validity of height and weight measurements, an appendix (Appendix E) on the reliability and validity of the four gospels, and an appendix (Appendix F) on the reliability and validity of claims regarding the effects of secondhand smoke. A list of references completes the work.

The book is replete with examples of various measurement situations (real and hypothetical), drawn from both the physical sciences and the social sciences. Measurement is at the heart of all sciences. Without reliable (and valid) instruments science would be impossible.

You may find my writing style to be a bit breezy. I can't help that; I write just like I talk (and nobody talks like some academics write!). I hope that my informal style has not led me to be any less rigorous in my arguments regarding the reliability of measuring measurements. If it has, I apologize to you and ask you to read no further if or when that happens. You may also feel that many of the references are old. Since I am a proponent of the "classical" approach to reliability, their inclusion is intentional.

I would like to thank Dr. Brown and Dr. Shlomo S. Sawilowsky (Wayne State University) for their very helpful comments regarding earlier manuscript versions

of the various chapters in this book. But don't hold them accountable for any mistakes that might remain. They're all mine.

Table of Contents

Chapter 5 The interpretation of individual measurements

Back to our hypothetical example, and a little more theory
How to interpret an individual measurement
- Point estimation
- Interval estimation
- Hypothesis testing

Compounded measurement error
Additional reading

Chapter 6 The reliability of difference scores

Types of difference scores
The general case
Measure-remeasure differences
Between-object differences
Change scores
- Simple change
- Controversy regarding the measurement of simple change
- Modified change
- Percent change
- Weighted change
- Residual change

Other difference scores that are not change scores
- Inter-instrument differences
- Inter- and intra-rater differences

Our flow meter example (revisited)
Additional reading

Chapter 7 The reliability of a single item

Single-item examples
X, T, and E for single dichotomous items
Some approaches to the estimation of the reliability of single items
- The Knapp method (and comparison to the phi coefficient)
- The Guttman method
- Percent agreement and Cohen's kappa
- Spearman-Brown in reverse

Visual analog(ue) scales
Additional reading

Chapter 8 The internal consistency of multi-item tests

A little history
Kuder and Richardson
Cronbach

CHAPTER 1: What do we mean by the reliability of a measuring instrument?

Contrary to its usual meaning in common parlance, in scientific measurement if something is reliable it is not necessarily good. (A security guard who falls asleep in the middle of his watch at exactly the same time every night would be a reliable sleeper, but he would undoubtedly be fired for his lack of alertness.) A reliable instrument is one that produces consistent measurements, which may or may not be worth anything. For example, a thermometer that reads 105 degrees Fahrenheit every time it is inserted in the mouth of a child who has no fever is reliable; it is consistent from one insertion to the next, but it is not a very good thermometer because it implies that the child has a high fever when (s)he actually does not. (It would be just as bad if it consistently yielded 98.6 for a febrile patient.)

Terminology

The matter of consistency of measurement is not referred to as “reliability” in all scientific disciplines. Some people prefer “accuracy”, “precision”, “agreement”, “dependability”, “reproducibility”, “repeatability”, or the term “consistency” itself. (See Goodenough, 1936, 1949; Stallings & Gillmore, 1971; Feinstein, 1985, 1987; Ennis, 1999; Norris, 1999; Last, 2001; and Dunn, 2004 regarding arguments for and against the use of some of those terms. Also see Gift & Soeken, 1988 and Lynn, 1989 for diametrically opposite notions of "accuracy", which is admittedly one of the most ambiguous terms.) The situation was so chaotic in the early part of the 20th century that Goodenough (1936) suggested doing away with the term “reliability” altogether and expressing any evidence regarding consistency of measurement strictly in terms of the procedures that were actually used to gather the evidence. There is a fairly recent article (Marks, Habicht, & Mueller, 1989) that even tried to distinguish among reliability, dependability, and precision, and the situation remains chaotic. I understand all of the arguments, but I still like "reliability".

There are also special “sub-kinds” of reliability, e.g., “test-retest” reliability, “parallel forms” reliability, “inter-rater” reliability, and many more. But all have in common some feature of consistency of measurement (from time to time, from item to item, from measurer to measurer, etc.)

To make things even more confusing, the term “reliability” is often used in the research literature in non-measurement contexts. In engineering and related disciplines, for example, equipment is said to be reliable if it doesn't break down, or if it is very unlikely to break down. (The engineering definition of reliability, as given by Blanchard, 1981, is: "The probability that a system or product will perform in a satisfactory manner for a given period of time when used under specified operating conditions.") And statisticians often use the term to refer to the sampling stability of a particular statistic such as a mean, a variance, or a

correlation coefficient. A value of a statistic, e.g., the proportion of Undecideds in a sample of 1000 likely voters prior to a presidential election, is said to be "reliable" if it is unlikely to fluctuate very much from one sample to another sample of the same size, i.e., its "margin of error" is small. I shall say nothing further in this book about reliability in the engineering sense. The sampling stability of various indicators of reliability will be considered in Chapter 11, but I (unlike Lincoln, 1932; 1933) will avoid using expressions such as "the unreliability of reliability coefficients"!

There are also many indicators of the degree of consistency, since no measuring instrument is perfectly reliable. Most of such indicators are correlation coefficients of some sort; others are variances or standard deviations; still others are statistics such as percentages or various functions of percentages.

Illustrative examples

A couple of examples would seem to be in order here. Consider first the thermometer referred to above. How might you determine to what extent it is a reliable measuring instrument? One way would be to insert it in the mouth of a particular individual several times (perhaps sterilizing it prior to each insertion), make a frequency distribution of the resulting measurements, and calculate the standard deviation of those measurements. The smaller the standard deviation, the more reliable the instrument (for that person on that occasion). Alternatively, you might measure each of several persons twice and calculate the correlation (the Pearson product-moment correlation coefficient, say) between the first and second readings and/or some other indicator of the association between corresponding (same person) first and second measurements.

Another example is the written essay examination. Is an essay question such as "What did you do on your summer vacation?" (a favorite of elementary school teachers) reliable? It may or may not be. In any event, one needs to clarify what kind of reliability is of concern. Test-retest? That is, would children asked to write an essay on that topic write essentially the same things in essentially the same ways if tested twice within, say, a 24-hour period? (The amount of time between test and retest is controversial--see, for example, my discussion of that matter in Knapp, 1985 and the further discussion in Knapp & Brown, 1995.) Or is it the reliability of the grading of a set of essays (written on that topic on a single occasion) by equally competent teachers that is important? If so, then some sort of inter-rater reliability evidence is essential. Or perhaps all that matters is whether a teacher who grades the essays agrees with (her)himself. In that case intra-rater reliability is of primary concern; that teacher must grade the essays twice, being "blinded" the second time to the names and the handwriting of the children, and with a sufficient amount of time passing between the two occasions so that (s)he doesn't recall what grade (s)he assigned the first time when doing the grading the second time. Gets complicated, doesn't it? (For more on the

matter of the reliability of the grading of essay examinations, see Gulliksen, 1936, 1950; Stanley, 1962; Coffman, 1972; and Braun, 1988).

Continuing with another educational testing example, how can you find out whether or not, or to what extent, a test of single-digit addition is reliable? Here the choices are even more numerous than for the essay-grading example. Suppose the test is to have 10 questions (items). Which of the following might you do?

a. Construct two "parallel" forms of the test by randomly sampling (without replacement) items of the type a+b=? from all of the possible permutations of such items (0+0=?; 0+1=?; …;9+9=?) to constitute the two forms, administer both forms to the same children, and compare the scores (total numbers of items answered correctly) on the two forms. This would give you some indication of how consistent the scores are from one version to another.
b. Construct just one form in that same manner, administer it twice to the same children, with a day or two in between testings, and compare the scores obtained on the two occasions. This is the "test-retest" approach and would tell you something about how stable the scores for that instrument are from one time to another.
c. Administer that one form once, and determine how consistent the children's performance is from one item to another, using one or more of the formulas for assessing the "internal consistency" of the instrument (see Chapter 8).
d. If scoring of the test is subject to individual judgment (for example, is that a 4 or a 9 that the child wrote?), get some evidence regarding either or both of inter-rater and intra-rater reliability.

[Note: One thing you might think of doing would be to administer one form of the single-digit addition test along with a test of general mathematical ability, and compare the scores on those two instruments. Unfortunately, that would tell you little or nothing about the reliability of the single-digit addition test. It might tell you something about its validity--see Appendix B.]

Necessity vs. sufficiency

These and similar matters are the "bread-and-butter" of what follows in the rest of this book. The mathematics is a bit heavy at times, but please do not lose the forest because of the trees that get in the way. Reliability is all about consistency, and nothing else. But consistency is a crucial property of a measuring instrument. To shift examples for illustrative purposes, a yardstick that consistently produces measurements of a person's height within an eighth of an inch of one another might not "really" measure height, but a yardstick that yields a different measurement every time it is used to measure a person's height is certainly undependable. In the language of mathematics, reliability is a necessary but not a sufficient condition for good measurement.

Additional reading

It is somewhat controversial whether you should refer to the reliability of a measuring instrument or the reliability of the measurements produced by the instrument. If it is clear that the determination of reliability is for this object (or these objects) on this occasion (or these occasions), no harm is done by referring to the result as an indicator of the reliability of the instrument itself, and that is my personal preference (thus the title of this book). But some people (e.g., Thompson, 1999, 2002, and elsewhere) get very bothered about so doing. For interesting exchanges on that topic, see Vache-Haase (1998), Sawilowsky (2000a; 2000b), Thompson and Vache-Haase (2000), a critique of Thompson (1999) by Sawilowsky and myself (Knapp & Sawilowsky, 2001a), his response (Thompson, 2001), and our rejoinder (Knapp & Sawilowsky, 2001b).

If you would like to read more about reliability in general, especially from a historical perspective, I recommend the articles by Kelley (1921, 1942), Cureton (1931, 1958), Cronbach (1947), Engstrom (1988), and Marradi (1990); Chapter XI on reliability in the book by McCall (1923); the section on reliability in my nursing research textbook (Knapp, 1998); the section on reliability and errors of measurement in the most recent "Standards" for educational and psychological tests (AERA, APA, & NCME, 1999); the chapters on reliability in the Educational Measurement compendia (Thorndike, 1951; Stanley, 1971, Feldt & Brennan, 1989; Haertel, 2006); and the digest by Rudner and Schafer (2001).

For a nice example of parallel-forms reliability, see the article by Beyer, Turner, et al. (2005) regarding the "Oucher" instrument for measuring children's self-reported pain. For some particularly good examples of the test-retest approach to reliability, see Grant, Hartford, et al. (1995), Grant, Dawson, et al. (2003), and Hasin, Carpenter, et al. (1997). For a very creative approach to inter-rater reliability for performance tasks where first ratings are given to the entire group of persons but second ratings are given to a randomly-chosen subgroup, see Livingston (2004). For further information regarding both the reliability and the validity of performance ratings, see Borman, Buck, et al. (2001).

Postscript: The title of this chapter is "What do we mean by the reliability of a measuring instrument?" Upon further reflection there is a prior question: "What do we mean by a measuring instrument?" In Chapter 14 and a couple of the later appendixes (Appendix E and Appendix F) I consider the reliability of claims. In that context the person making a claim is "the measuring instrument". Does that make sense. After you read Chapter 14 and those two appendixes, please e-mail me at tknapp5@juno.com and let me know what you think. Thanks.

CHAPTER 2: Measurement error

Before discussing the matter of measurement error, I would like to clarify the difference between reliability and validity.

Attribute vs. variable

It is essential to make the crucial distinction between the attribute we are trying to measure (e.g., intelligence) and the variable that we are directly concerned with (e.g., score on the Wechsler Adult Intelligence Scale). The former is an abstract entity (psychologists call such things "constructs"), while the latter is an attempt to "concretize" (or "operationalize") the abstraction. God only knows what Mary Smith's intelligence is. We mere mortals must settle for trying to get a good fix on the measurement for Mary on some instrument such as the Wechsler Adult Intelligence Scale. The "fit" between the attribute and the variable is a matter of measurement validity, which as I pointed out in the Preface ultimately comes down to a matter of expert judgment. The "fit" between the measurement actually obtained on a variable and the measurement that in some sense should have been obtained on that variable (the so-called "true score", which God alone also knows--see below) is a matter of measurement reliability (Knapp, 1985). The latter is the focus of this book. (See Symonds, 1928; Adams, 1936; Cureton, 1950, 1965; Cattell, 1964; Heise & Bohrnstedt, 1970; Kaiser & Carter, 1971; Terwilliger & Lele, 1979; and Suen, 1987 for interesting distinctions and connections among reliability, validity, and other measurement terms such as "homogeneity", "objectivity" and "relevance".)

At the end of the previous chapter I said that reliability is a necessary but not sufficient condition for good measurement. "Good measurement" encompasses a lot of things: reliability, validity, objectivity, usability, etc. The fascinating question is whether reliability is a necessary condition for validity or if an instrument can be valid yet not reliable. That matter has been extrensively debated for many years, with the most recent debate taking place in the pages of the Educational Researcher and the Journal of Educational and Behavioral Statistics. Moss (1994) claimed that you can have validity without reliability. Li (2003) and Mislevy (2004) disagreed; they provided both theoretical and practical reasons, while also appealing to the arguments given by Linn (1994) in his discussion of recent educational assessment instruments. In her rejoinder to Mislevy, Moss (2004) remained firm in her claim. It is a very complicated problem, with the controversy spilling over into considerations regarding "randomness" (see following section), theory vs. practice, and the distinctions between "qualitative" and "quantitative" research. I urge all of you who are interested in educational research to read those articles and draw your own conclusions.

When is something random?

A discussion of randomness is now in order. In many measurement textbooks the distinction is drawn between “constant” errors of measurement (e.g., being always or almost always off on the high side by a couple of inches when measuring height) and what some people call “random” errors of measurement (e.g., being sometimes off by varying amounts on the high side and being sometimes off by varying amounts on the low side). The former kinds of errors are usually ascribed to invalidity and the latter kinds to unreliability. But how can one determine whether or not an error of measurement is random? Aye, there's the rub. Volumes have been written on the concept of randomness (random sampling, random assignment, random ordering, etc.), but there appears to be no consensus regarding whether randomness pertains to the process or to the product of some endeavor. Many people (Wallis & Roberts, 1962 being among them) would argue, for example, that if a deck of cards is properly shuffled, 13 cards are drawn, and they're all spades, that sample is a random (though unlikely and surprising) sample. Other people (e.g., Siegel & Castellan, 1988) would argue that some sort of "test of randomness" should be made, and "passed", before a sample is declared to be a random sample.

What is the relevance of this for measurement reliability? The principal (and happiest) consequence is that all of the theorems, formulas, etc. of "classical" reliability theory can be derived with and without the assumption of randomness (process or product), and this will be shown in the following chapter. The arguments all revolve around the concepts of true score and error score, to which I shall now turn.

Obtained score, true score, and error score

Consider a spelling test that consists of 50 words randomly drawn (the process being random) from an unabridged dictionary, dictated to a group of examinees, with each examinee asked to write down the spelling of each word (as it is called out by an examiner) on a sheet of paper containing the numbers from 1 to 50. The score on the test is the number of words spelled correctly. For each of the examinees we can envision three scores: (1) the number of words (s)he spells correctly; (2) the number of words (s)he should have been able to spell correctly; and (3) some function of the discrepancy between those two numbers. In the classical measurement theory developed by psychologists and educators (see, for example, Gulliksen, 1950; Lord & Novick, 1968), the first of these is called the individual's "obtained score" or "observed score"; the second is called (her)his "true score" (Cronbach, 1970 and elsewhere--and a few others prefer the term "universe score"); and the simple difference (obtained minus true) is called an "error score". (The number of words spelled incorrectly is also an obtained score, even though it is a count of the number of spelling errors. Do you follow those two different meanings of “error”?)

The obtained score might be higher than the corresponding true score (if, for example, the person "got lucky" in the sense that the specific 50 words on the test were among those that (s)he knew best how to spell). Or the obtained score might be lower than the true score (if, for example, probability dealt the person an unlucky blow and those 50 words were among (her)his "sticklers"). Or, but most unlikely, the obtained score might be identical to the true score, and there would be no error. The problem, of course, is that all we can know are the obtained scores, and we can only speculate about the corresponding true and error scores. Such speculation usually involves assumptions concerning true scores and error scores, a matter to be considered in great length in the following chapter.

Dunn's example

Dunn (2004) starts out his book on measurement error with a very interesting example of two attempts--with and without "zeroing" between measurements--to determine the reliability of a kitchen scale based upon repeated measurements of a packet of dried fruit alleged to weigh 500 grams (it says so on the package). The word "alleged" is important. Some people argue that the true weight of an object, e.g., a one-pound ball at the National Institute of Standards and Technology (NIST), is often known. Not so. Every measuring instrument, including the instrument that designated the "one-pound" ball for the NIST in the first place, no matter how "accurate" it is said to be, is imperfect and is subject to measurement error, no matter how small. Again, God only knows the true weight of that one-pound ball.

[Exercise for the reader: Read the article by Greenland, Bowley, et al. (1990) very carefully and see if you agree with their arguments regarding "true" cholesterol values.]

Continuous vs. discrete variables

The previous definitions assume that the measurements (e.g., number of words spelled correctly) are continuous or "continuous enough" so that the matters of addition and subtraction are defensible. For other kinds of measurements, particularly dichotomous variables that can take on just two discrete values, a and b, and may even be categorical (e.g., a = "yes" and b = "no"), obtained score, true score, and error score must be formulated differently (see Chapter 7). Things can get even more complicated for variables that are ordinal scales (see Chapter 10).

The controversial true score

Do these concepts (obtained score, true score, error score) make sense to you? They do to most measurement theorists (especially Lord, 1959c), but some, e.g., Loevinger (1957), Ross and Lumsden (1968)--see also Lumsden's (1976)

review of test theory--and Cliff (1979) think that we should do away with the concept of true score (and with most if not all of reliability theory), concentrating entirely on the validity of measuring instruments and their associated obtained scores only. I happen to disagree with them (if I agreed with them I wouldn't have attempted to write a whole book on reliability!), but I urge those of you who might be interested to read their articles very carefully--after you've finished reading this book--and decide for yourselves whether or not the concept of true score should be abandoned.

Is true score a latent variable? There has been a great deal written about that in the measurement literature. Schmidt and Hunter (1999) claimed that you should consider a true score as both unknown and latent (underlying an obtained score). In their commentary regarding Schmidt and Hunter's claim, Borsboom and Mellenbergh (2002) claimed that true score, though unknown, is not latent. They argued that any considerations concerning latency were matters of validity, not reliability. Other authors took less extreme positions with respect to the question. My colleague, Hak P. Tam, and I have attempted to summarize the various arguments (Knapp & Tam, 2007). It appears to depend upon how you define the term "latent variable", as Bollen (2002) had so carefully pointed out.

Some more thoughts about randomness

Returning to the matter of measurement errors, when are they "random"? For the spelling test example they do indeed appear to be random, because chance and chance alone determined which words would be on the test, and any discrepancy between a person's obtained score and (her)his true score is an error that occurs by chance. But for other equally important measurement situations, e.g., measuring a person's height, I would be hard-pressed to call the difference between a person's obtained height and (her)his true height a "random" error. It would seem likely that something other than chance produced an obtained height of 60 inches, say, if the person's true height were 64 inches (an error of four inches). What might have happened? If a stadiometer (that part of the scale in doctors' offices that measures height) was used, the person may not have been standing up straight; or the healthcare provider may have read off the height incorrectly; or whatever. But those eventualities aren't random, are they? Or are they? Some, I dare say most, measurement experts (see, for example, McCall, 1923; Suen, 1990) are willing to call such errors random; I am not (I favor the argument that randomness is relevant to a process and not a product). Fortunately, it doesn't matter which "philosophical" stance you take on this matter. Obtained scores, true scores, and error scores can be formulated with or without the assumption of random errors of measurement, which Gulliksen (1950) so clearly demonstrated in the second and third chapters of his text and which I will try to summarize in the chapter that follows.

Accuracy, bias, scale differential, and precision

As I indicated in the previous chapter, the term "accuracy" means different things to different people. The most common use has to do with validity; an instrument is called "accurate" if it measures correctly what you want it to measure. Some authors (e.g., van Belle, 2002) equate accuracy with lack of "bias" (where "bias" is defined as the difference between the location of obtained measurements for the instrument in question and the location of "gold standard" measurements). Some of those same authors (see, for example, Lin, 1989 and Liao, 2003) have suggested various indexes (called concordance correlation coefficients) that summarize for a given instrument its bias, "scale differential" (the discrepancy between the instrument's and the gold standard's variances), and "precision" (which they take to be synonymous with reliability). I personally shy away from all of those terms, and I agree with Feinstein (1985) regarding his arguments against their use when referring to the consistency of a measuring instrument.

Additional reading

For more on measurement error, I recommend Cureton (1931), Grubbs (1948; 1973), Cochran (1968), Hanamura (1975), Topping (1975), Cameron (1982), Jaech (1985), Becker (2000), and Schmidt, Le, and Ilies (2003). [Becker's article and Schmidt, et al.'s article are devoted to the determination of the effect of "transient error" on reliability, i.e., measurement error attributable to momentary, unrepeatable contexts.] For interesting discussions of measurement error associated with using a surrogate measuring instrument rather than a theoretically more appropriate but unavailable instrument, including how to adjust for surrogacy, see Gustafson (2004) and some of the references cited in his book, especially Bashir and Duffy (1997); Bashir, Duffy, and Qizilbach (1997); Brown, Krieger, et al. (2001); Rosner and Gore (2001); and Zidek, Wong, et al. (1996).

CHAPTER 3: Reliability theory (abridged, with examples)

The "classical" theory of reliability proceeds as follows.

The basic concepts

Let X = an obtained measurement; let T = the associated true measurement; and let E = the associated measurement error. In the educational and psychological literature, where most of reliability theory has been derived, the terms "obtained score", "true score", and "error score" are used. (See previous chapter.) In other scientific literature the concept of a "score" is often not relevant, but I will continue to use the score terminology to apply to any measurement context where X is what we know, and T and E are what we wished we knew.

The first few axioms, definitions, and theorems

Axiom #1 (unprovable but reasonable assumption): $X = T + E$.

This axiom proposes that any obtained score is the simple sum of the associated true and error scores. Therefore we also have $T = X - E$ and $E = X - T$. Let us first define error score, and true score will fall out as the difference between X and E. Later on in the chapter we will take the opposite approach by first defining true score and letting error score fall out as the difference between X and T.

Definition #1: An error score, E, is a number in the same metric as X (both are in inches, for example), with the properties (a) it has a mean of 0 when several objects are measured with the instrument; and (b) it is uncorrelated with everything that is unknown--true score for that same instrument, error score for any other instrument, etc.

Those two properties are what are alleged to make an error "random" (in the product sense). Having a mean of 0 implies that there is no systematic bias; for some objects you're off on the high side and for others you're off on the low side. Being uncorrelated with true score and other error scores implies that it "plays no favorites"; a positive error is just as likely to be associated with a low true score as a high true score, for example.

Theorem #1: The mean of X is equal to the mean of T, i.e., $M_X = M_T$.

Proof: Since $X = T+E$, $M_X = M_{T+E}$. But $M_{T+E} = M_T + M_E$ (from basic statistics) and $M_E = 0$ (from the preceding definition of error score). Therefore $M_X = M_T$.

This theorem (if satisfied) indicates that no matter how reliable or unreliable an instrument may be, when you measure several objects with it, the mean of the obtained scores that you do get is equal to the mean of the true scores that you would have gotten (if they were known). That's not saying much, but it's a start.

Theorem #2: The variance of the obtained scores, X, is equal to the variance of the true scores,T, plus the variance of the error scores, E, i.e., $S_X^2 = S_T^2 + S_E^2$.

Proof: It can be shown in mathematical statistics that the variance of any sum is equal to the sum of the corresponding variances plus twice the sum of the pairwise covariances. Since X = T + E, the variance of X is equal to the variance of T plus the variance of E plus twice the covariance of T and E. It can further be shown that the covariance of T and E is the correlation (Pearson product-moment) between T and E multiplied by the product of the standard deviation of T and the standard deviation of E. But the correlation between T and E is equal to zero by the above definition of an error score. Therefore, the variance of X reduces to the sum of the variance of T and the variance of E.

Note that the standard deviation of X is NOT equal to the standard deviation of T plus the standard deviation of E. Theorem #2 is just like the Pythagorean theorem; $c^2 = a^2 + b^2$, but $c \neq a + b$. The standard deviation of E, i.e., the standard deviation of the errors of measurement (the square root of the variance of E), is given a special name, the standard error of measurement, and it will be afforded considerable attention later in this chapter and in Chapter 5.

This second theorem indicates that the dispersion (spread) of obtained scores is attributed to the dispersion of the true scores and the extent to which the dispersion of those true scores has been "inflated" by measurement error. If all of the obtained variance is true variance (unlikely), then there is no error variance. On the other hand, if all of the obtained variance is error variance (also unlikely), then there is no true variance (all of the objects being measured have the same true score but get different obtained scores solely because of measurement error). One would expect that for most measurement situations the true variance would represent most, but not all, of the total obtained variance.

Definition #2: The proportion of obtained variance that is true variance, i.e., S_T^2/S_X^2, is a quantity (although unknown, since S_T^2 is unknown) we shall refer to as the reliability coefficient, r_{XX}.

The reliability coefficient is not a coefficient in the strict mathematical sense of that word (it doesn't necessarily multiply anything) and it is given the symbol r_{XX} because it is sort of a "self-correlation" of X with X--see Cronbach (1947) and Horst (1954). It can take on any value between 0 and 1, and there is a vast literature on various ways to estimate it. I shall discuss much of that literature throughout the rest of this book.

A hypothetical example

Suppose that you measure basketball player Mary Smith's height once with a particular yardstick manufactured by the John Jones Company. (Have her stand against a wall with her head in the so-called "Frankfort plane", then ask her to step away and permit you to measure from the floor to the spot on the wall corresponding to the top of her head. If she is more than three feet tall you'll have to slide the yardstick along once or twice.) You determine that her obtained score, X, is 60 inches, i.e., 5 feet, no inches. Her unknown true score, T, is in the same metric as X; so is her unknown error score, E. Any combination of T and E could have produced X (59 + 1, 62 - 2, 60 + 0, etc.). Moreover, that particular yardstick could be a poor way to measure height (but that's validity, not reliability; please don't ever forget that). Assume that Mary's error score is -4 inches (God tells you that), so that her true score is 64 inches (by solving for T in the equation T = X - E).

Now measure the heights of six other basketball players with that same yardstick. (Six more players won't give us "several" people, but it will suffice for our present purposes.) Suppose that you obtain the following heights X, God provides the errors E, and you calculate the true scores T. These data are purely hypothetical, but they satisfy all of the previous definitions, axioms, and theorems (check that for yourself or see below).

Person	X	T	E
1 (Mary)	60	64	-4
2 (Carol)	64	66	-2
3 (Alice)	74	68	+6
4 (Bob)	72	70	+2
5 (Ted)	72	72	0
6 (Joe)	78	74	+4
7 (Jim)	70	76	-6

A different approach

And now, the promised alternative approach to X, T, and E.

Definition #3: An object's true score, T, is the average of the object's obtained parallel measurements, X, for a very large number of such measurements.

Talk about a theoretical notion! What do we mean by "parallel" measurements? How large is very large? It's tough enough to measure an object once; how can we measure an object more than once without affecting its true score?

Those are all very good questions. Let's take them one at a time:

Definition #4: Two or more forms of a measuring instrument are called parallel (equivalent, alternative, exchangeable, comparable) if an object's true score is the same for all forms (a judgment call) and if all forms produce equal means, equal variances, and equal inter-correlations when applied to a very large number of objects at the same point in time.

That helps. You would be hard-pressed, for example, to consider two thermometers to be parallel if there were systematic differences between the temperature readings produced when they were used to measure the same persons at approximately the same time. And you would also not be willing to regard two history tests to be parallel if one of them were much more difficult than the other for all examinees. But this leaves the matter of "very large" still unresolved. Mathematicians like to talk about "approaching infinity", but that's not very comforting; let's leave it at "very large", at least for the time being.

Measuring each object more than once remains a practical problem in some situations (e.g., many temperature measurements taken on many persons), but not others (e.g., many spelling items administered to many persons). In any event, we shall take the perhaps-debatable position that an object's true score remains constant, and is unaffected by the measurement process itself, at least during the particular time period in which the instrument is to be used. (See Hoffman, 1963 regarding some alterations to reliability theory when true scores might change due to practice effects.)

The easy part now is that if we still assume that X = T+E, E falls out by simple subtraction of T from X, i.e., E = X -T. Unfortunately, however, the two theorems stated above ($M_X = M_T + M_E$ and $S_X^2 = S_T^2 + S_E^2$) cannot be proven in the same way, because those proofs depended upon the concept of random error and its associated properties. But the reliability coefficient will continue to be defined as S_T^2 / S_X^2.

Alternative proof of Theorem #1: $M_X = M_T$

By the new definition of true score, T for any object is the average across parallel forms of the corresponding X's for that same object. Therefore, the average T across objects is the same as the average across objects of the average X for each object. But that is the same as the average X; hence $M_T = M_X$, or, equivalently, $M_X = M_T$.

Proof that $M_E = 0$:

$M_E = M_{X-T}$ (since E=X -T). $M_{X-T} = M_X - M_T$ (from basic statistics). But $M_X = M_T$. Therefore, $M_E = 0$.
[This was part of the definition of error score earlier on, but it needed to be proven here.]

Proof that $r_{TE} = 0$:

The correlation (Pearson product-moment) between any two variables U and V can be written as $r_{UV} = (\sum UV - NM_UM_V) / NS_US_V$, where N is the number of objects for which you have data, the M's are the means, and the S's are the standard deviations (trust me).

Letting U = T and V = E, we have

$r_{TE} = (\sum TE - NM_TM_E)/ NS_TS_E = \{\sum T(X -T) - NM_TM_{X-T})\} / NS_TS_{X-T}$

Expanding, simplifying, and appealing to Theorem #1, we have

$r_{TE} = \{\sum TX - \sum T^2 - NM_T(M_T - M_T)\} / NS_TS_{X-T}$

But $\sum TX$ is actually a double summation across objects and parallel forms of X and is equal to $\sum T^2$. Therefore the numerator of the expression for r_{TE} is equal to 0 and r_{TE} itself is equal to 0.

[This was also part of the definition of error score if errors are assumed to be random, but it also needed to be proven here.]

Alternative proof of Theorem #2: $S_X^2 = S_T^2 + S_E^2$.

The variance of the error scores, S_E^2, is equal to S_{X-T}^2 (since E=X -T), which from mathematical statistics is equal to the sum of the variance of X and the variance of T minus twice the covariance of X and T, that covariance being equal to the product of the correlation between X and T, the standard deviation of X, and the standard deviation of T; i.e.,

$S_E^2 = S_X^2 + S_T^2 - 2r_{XT} S_XS_T$.

If $r_{XT} S_X$ can be shown to be equal to S_T the theorem will be proven, since the right-hand side of the equation would reduce to $S_X^2 - S_T^2$ and $S_E^2 = S_X^2 - S_T^2$ is the same as $S_X^2 = S_T^2 + S_E^2$. That's a bit tricky, but here goes.

$r_{XT} = (\sum XT - NM_T^2) /NS_XS_T$ from the general formula for a Pearson r. But $\sum XT = \sum TX$ is a double summation across objects and forms (see above) and $\sum TX = \sum T^2$. Substituting $\sum T^2$ for $\sum XT$ in the formula for r_{XT} and re-arranging slightly, we have that $r_{XT} = \{(\sum T^2 - NM_T^2)/N\} /S_X S_T$. But the term inside the braces is equal to S_T^2, because the variance of any variable U can be written in the form $(\sum U^2 - NM_U^2)/N$ (trust me regarding that also), giving $r_{XT} = S_T/S_X$ or $r_{XT} S_X = S_T$, as advertised.

Since true scores are assumed to be constant across parallel forms for which obtained variances are defined to be equal, the variance of the true scores is

also constant, and a consequence of this theorem is that the error variances of the forms are also equal, differing from the obtained variance by that constant.

[I hate to have to say it, but some people (e.g., Votaw, 1948) include in the definition of parallel forms the requirement that the forms also have to have equal validity for predicting some external criterion--see the discussion of criterion-related validity in Appendix B--but other people (like me) do not, because they feel that validity is a separate matter.]

Some other concepts and terminology

The expression for the correlation between obtained scores and true scores ($r_{XT} = S_T/S_X$) is the ratio of the standard deviation of the true scores to the standard deviation of the obtained scores, which is in turn equal to the square root of the reliability coefficient, i.e., $r_{XT} = \sqrt{r_{XX}}$, and it is often referred to as the "index of reliability". Some authors of measurement textbooks prefer to emphasize its square, r_{XT}^2 , which is equal to $S_T^2 / S_X^2 = r_{XX}$, so that it can be interpreted just like any other squared correlation as the proportion of the variance of X that is "accounted for" by the variance of T. [Some even square the r_{XX} ; that's just plain wrong.]

An interesting variation of the reliability coeficient is the signal-to-noise ratio, which is equal to $r_{XX} / (1 - r_{XX})$, for $r_{XX} \neq 1$. See Cronbach and Gleser (1964) regarding some advantages and some disadvantages of that ratio.

In the educational and psychological measurement literature two forms of a measuring instrument are called "tau-equivalent" if each object's true score is the same for both forms but the error variances are not equal. They're called "essentially tau equivalent" if the true scores differ only by an additive constant and the error variances are unequal. And they're called "congeneric" if the true score on any one form is a linear transformation of the true score on any other form, again for unequal error variances. See Traub (1994), Dunn (2004), or Graham (2006) for a discussion of those variations on parallelism. They will not be treated any further in this book.

Topping (1975) allows for systematic bias by defining "accuracy" in terms of the difference between T and the average X for a given object, and by defining "precision" in terms of the variance of the X's from one another regardless of the corresponding T for that object. Chalk up another instance of terminological confusion?

The key theorem

Although we seem to be making some progress in being able to prove a couple of theorems, we still have too many unknowns. How can we convert at least some of those unknowns into knowns? One very promising way is to see if we can prove the following theorem.

Theorem #3: The reliability coefficient, r_{XX}, for a particular instrument, is equal to the correlation between parallel forms, A and B, of that instrument; i.e., $r_{XX} = r_{AB}$.

Proof: The correlation between A and B can be written as $(\sum AB - NM_AM_B)/NS_AS_B$ --see above. But $M_A = M_B = M_T$, by the definition of parallel forms and by Theorem #1. Therefore, this expression can be written as $(\sum AB - NM_T^2) /NS_AS_B$. Furthermore, $\sum AB = \sum(T+E_A)(T+E_B) = \sum (T^2 + TE_A + TE_B + E_AE_B)$. Summing those individual terms we have $\sum T^2 + \sum TE_A + \sum TE_B + \sum E_AE_B$. The two middle terms are equal to zero. (That is because true and error scores were defined to be uncorrelated according to the concept of random error, and then proven to be uncorrelated without appealing to that concept. Since those correlations are equal to zero so are the sums of the corresponding cross-products, whenever M_E also is equal to zero.). That leaves $\sum T^2$ and $\sum E_AE_B$. The latter term is equal to zero if we assume the E's are random. But if we don't, we have to prove that it is equal to zero. Here goes.

$\sum E_AE_B = \sum(X_A - T)(X_B - T) = \sum X_AX_B - \sum X_AT - \sum X_BT + \sum T^2$ [expanding and summing]

But by the above argument concerning double summation, all four of those summations can be written as $\sum T^2$. Since two of them are positive and two of them are negative, $\sum E_AE_B = 0$.

Returning to the expression for the correlation between two parallel forms, it can now be written as $\sum(T^2 - NM_T^2) /NS_X^2$, because $S_A = S_B = S_X$ (A and B being parallel forms of X). That expression in turn is the product of S_T^2 (see above) and $1/S_X^2$. Therefore the correlation between two parallel forms is equal to S_T^2/S_X^2, or r_{XX}, the reliability coefficient.

Whew! But all of that algebra was worth it, because we now have a way of calculating (or at least estimating) a reliability coefficient: Construct, or try to construct, two parallel forms of an instrument, get obtained scores on both forms for a very large number of objects, and correlate the two. And you don't have to worry about whether or not the measurement errors are random if you are willing to accept the definition of a true score as the (unknown) average of parallel measurements. An additional benefit is that we can also calculate (estimate) the standard error of measurement. Since $S_X^2 = S_T^2 + S_E^2$ by Theorem #2, and $r_{AB} = r_{XX} = S_T^2/S_X^2$ by Theorem #3, re-solving the first expression for S_E^2, we have

$S_E^2 = S_X^2 - S_T^2 = S_X^2 - r_{AB} S_X^2 = S_X^2 (1 - r_{AB})$, so that $S_E = S_X \sqrt{(1 - r_{AB})}$.

A caution concerning parallelism and reliability

It should go without saying, but I'll say it anyhow, that parallelism does not guarantee reliability. You could have two perfectly parallel forms whose measurements correlate zero with one another, and the forms would therefore be perfectly unreliable. This is the same situation that could prevail if the means and variances for two identically-scaled instruments, e.g., a math test and an English test, were equal, but there was a zero correlation between them. Equal means and equal variances tell you nothing about how related two variables are.

Truman Kelley on parallelism and reliability

Truman L. Kelley (1921, 1923, 1925, 1927, 1942, 1947, and many other sources) was one of the pioneers in educational measurement and statistics. He took an interesting and controversial stance on parallelism and reliability. First of all, he insisted that the parallel forms approach (full forms or half forms--see Chapter 8) was the only defensible approach to reliability. (Louis Guttman, 1945 and elsewhere, another authority on measurement and statistics, claimed that test-retest was the proper approach; but see Hoffman, 1963 for the problem of practice effects.) Secondly, he (Kelley) argued that the determination of parallelism was primarily an act of judgment and (unlike what I have postulated in this chapter--see above definition of parallelism) not a mathematical matter. Finally, he defined the reliability coefficient as the correlation between all possible inter-individual differences on one form of an instrument and all possible inter-individual differences on a parallel form, which he showed (Kelley, 1942) to be equal to the correlation between the obtained scores on the two forms and, by transitivity, to be also equal to the ratio of true variance to obtained variance. He rejected the test-retest approach because he claimed that the mental processes at Time 2 are different from those at Time 1. And he referred to all of the internal consistency measures (see Chapter 8) as "cohesion" or "coherence" coefficients, not reliability coefficients. His arguments are very persuasive, but unfortunately (for Kelley) he lost the fight. (See the discussion by Traub, 1997, of that fascinating controversy.)

A much-later reflection of Kelley's approach was the postulation by Savedra, et al. (1989) of the parallelism of two forms for pain identification by children: (1) marking on body outlines; and (2) and pointing to locations on their own bodies.

A couple of examples (one hypothetical, one real)

The first example is an extension of the example of the heights of seven basketball players treated earlier in this chapter. Suppose that you measure those same people a second time with another yardstick manufactured by the John Jones Company, and God provides you with the corresponding set of error

scores (the true scores are assumed to remain the same). Here are the data; the first, third, and fourth columns of data are the same as the three previous columns of data. The measurements are all in inches.

Hypothetical data

Person	First X	Second X	True T	First E	Second E
1 (Mary)	60	64	64	-4	0
2 (Carol)	64	64	66	-2	-2
3 (Alice)	74	74	68	+6	+6
4 (Bob)	72	64	70	+2	-6
5 (Ted)	72	76	72	0	+4
6 (Joe)	78	70	74	+4	-4
7 (Jim)	70	78	76	-6	+2

By merely "eyeballing" these data it would appear that the yardsticks are unreliable, since those true heights in general aren't very close to the corresponding obtained heights, and the errors are all over the place. But just how unreliable are they (for these seven people on these two occasions)?

First, some basic statistics:

$N = 7$

$M_{X1} = 70$ (the mean of the first set of obtained heights)

$M_{X2} = 70$ (the mean of the second set of obtained heights)

$S_{X1}^2 = 32$ (the variance of the first set of obtained heights)

$S_{X2}^2 = 32$ (the variance of the second set of obtained heights)

[Note that both of these variances were calculated by dividing the sum of the squared deviations by N. You only use N-1 when you have a random sample and you want to get an unbiased estimate of a population variance--see Knapp, 1970. I am treating these seven people as a population.]

That's comforting; the two yardsticks appear to be "parallel". Let's call both of those means M_X and both of those variances S_X^2, and calculate some more reliability-relevant statistics.

$M_T = 70$

Theorem #1 is satisfied.

$M_{E1} = M_{E2} = 0 = M_E$
$r_{TE1} = r_{TE2} = r_{E1E2} = 0$

That's very nice. The errors appear to be "random".

$S_T^2 = 16$

$S_{E1}^2 = S_{E2}^2 = 16 = S_E^2$ ($S_E = \sqrt{16} = 4$)

$S_X^2 = 32 = S_T^2 + S_E^2 = 16 + 16$

Theorem #2 is satisfied.

$r_{X1X2} = r_{AB} = .50 = S_T^2/S_X^2 = r_{XX}$ (the reliability coefficient). Also $r_{X1T} = r_{X2T} = r_{XT} = \sqrt{r_{XX}} = .71$

Theorem #3 is satisfied.

S_E can also be determined by calculating $S_X \sqrt{(1 - r_{XX})} = S_X \sqrt{(1 - r_{AB})} = \sqrt{32}$ times $\sqrt{(1 - .50)} = 4$ (see above). And there is even a third formula for S_E, which is particularly useful for estimating the standard error of measurement for real data (where all of the assumptions underlying classical reliability theory may not be perfectly satisfied), and it is $\sqrt{(\sum d_i^2 / 2N)}$, where d_i is the difference between paired simultaneously-obtained measurements for object i and N is the total number of objects being measured (i = 1,2,...,N). In some of the measurement literature the quantity yielded by this formula is called the "technical error of measurement" (TEM), rather than the standard error of measurement. For our hypothetical example, the d_i (= First X - Second X) are -4, 0, 0, 8, -4, 8, and -8, respectively; their squares are 16, 0, 0, 64, 16, 64, and 64, again respectively; the sum of those squares is 224; and 2N is 14. Substituting the last two numbers in the formula for TEM, we have $\sqrt{(224/14)} = \sqrt{16} = 4$. For more on the technical error of measurement, see the articles by Engstrom (1988) and by DeKeyser and Pugh (1990); and my article on TEM in the American Journal of Physical Anthropology--Knapp, 1992 (along with the references cited therein).

[Exercise for the reader: I have not given you the correlations between the X's and the E's. If you have easy access to a statistical computer package such as MINITAB or SPSS or SAS, enter the above data into your computer and ask your package to calculate all ten of the pairwise correlations for the five columns of numbers and see what you get. You may be in for a surprise. It turns out that the correlation between X (either X) and E (either E) is equal to $\sqrt{(1 - r_{XX})}$, which is equal to zero only if $r_{XX} = 1$ (for these data it just happens that r_{XE} is the same as r_{XT}, i.e., .71), whereas the correlation between T and E (either E) is always equal to zero in classical reliability theory.]

Regarding the question posed above (how unreliable are the Jones yardsticks?), based upon these data (which are admittedly meager, but that's all we have!), the answer is "very", since only 50% of the variation in the obtained measurements is "accounted for" by the variation in the true measurements. The other half of the variation can be attributed to measurement error. Would you like to use either of those yardsticks to measure people's heights? I wouldn't.

Real data

In real life we will not have perfectly parallel forms of our measuring instruments, and we'll never know the true scores and the error scores, so we'll have to make some additional assumptions and be willing to settle for all sorts of approximations to the "real" reliabilities of those instruments. Consider, for example, the data obtained by Bland and Altman (1986) in their investigation of the measure-remeasure reliabilities of two instruments for measuring the peak expiratory flow rate (PEFR) of 17 subjects. (See also Altman & Bland, 1983; Bland & Altman, 1999. They and a few other authors, e.g., Haber & Barnhart, 2008, and Yi, Wang, & He, 2008, are interested in "method comparison". It is a term that is not usually concerned specifically with reliability or validity, but refers to an investigation of both within-instrument and between-instrument agreement, with neither taken as a gold standard for the other.) The instruments were a Wright Peak Flow Meter and a "Mini" Wright Peak Flow Meter, and the obtained data (expressed in liters per minute) were as follows, using X for Wright and Y for Mini Wright:

	Wright		Mini Wright	
Subject	First X	Second X	First Y	Second Y
1	494	490	512	525
2	395	397	430	415
3	516	512	520	508
4	434	401	428	444
5	476	500	500	500
6	557	611	600	625
7	413	415	364	460
8	442	432	380	390
9	650	638	658	642
10	433	429	445	432
11	417	420	432	420
12	656	633	626	605
13	267	275	260	227
14	478	492	477	467
15	178	165	259	268
16	423	372	350	370
17	427	421	451	443

[Note that I use "object" and "subject" interchangeably if the objects are people, contrary to the use of those two terms in English grammar, where "object" and "subject" are definitely not interchangeable!]

These data show that the two measurement occasions weren't exactly parallel for either instrument. The mean First X for the Wright data is 450.4, whereas the mean Second X for the Wright data is 445.4 (the variances are 12730.06 and 13462.74, respectively). The mean First Y for the Mini Wright data is 452.5, whereas the mean Second Y for the Mini Wright data is 455.4 (the variances are 12039.16 and 11659.00, respectively). But both instruments appear to be very reliable, with the correlation between First X and Second X for Wright = .983 and with the correlation between First Y and Second Y for Mini Wright = .967. The other correlations are .943 between First X and First Y, .936 between First X and Second Y, .957 between Second X and First Y, and .952 between Second X and Second Y. Those are surprisingly high, aren't they, especially the correlation between First X and Second X and between First Y and Second Y, given, for example, the discrepancies between the First X and the Second X for Subject 6 and for Subject 16 on the Wright, and the discrepancy between the First Y and the Second Y for Subject 7 on the Mini Wright? But keep in mind that correlation coefficients are indicative of relative, not absolute, agreement--Bland and Altman (1986) argue that point very well--and those correlations are only equal to the reliability coefficients when all of the tenets of classical reliability theory are satisfied.

The standard error of measurement for both instruments is approximately 15 to 20 liters per minute. This is a bit difficult to estimate, since it is theoretically equal to $S_X \sqrt{(1 - r_{XX})}$, and the standard deviations for the obtained measurements on the two occasions are not the same for either instrument, but if we average the two variances for the Wright data, take the square root of that average variance, and use that for S_X in the formula for the standard error of measurement, we get 14.9. (In a similar manner we get 19.9 for the Mini Wright data). I will explain how a standard error of measurement should be used when I discuss the interpretation of individual measurements in Chapter 5.

[Another exercise for the reader who has easy access to a computer and a statistical package: Enter the Bland & Altman data, ask your software to calculate all of the means, variances, and correlations, and see if you get the same answers that I did, at least to the same number of decimal places. (I have argued elsewhere--Knapp, 2001--that you need not report reliability evidence to all of those decimal places that are provided by most modern statistical packages.) You might also want to make a couple of scatterplots, e.g., Second X against First X for the Wright meter, to see if the relationship between the two variables looks linear.]

Additional reading

As indicated by the title of this chapter, the foregoing has been an abridged (and simplified) account of the basic concepts in classical reliability theory. If you would like to read a similar abridged account, see Chapter 5 of Furr and Bacharach (2008) and/or Chapter 5 of my little paperback book on statistics for educational measurement (Knapp, 1971)--with all its warts (lots of typos, so be careful). If you would like to pursue such theory in greater depth, I strongly recommend that you read Gulliksen's (1950) "old" but excellent text, especially the first three chapters. You might also want to look up the chapters by Thorndike (1951), Stanley (1971), Feldt and Brennan (1989), and Haertel (2006) in the Educational Measurement compendia; the chapter by Bohrnstedt (1983) in the Handbook of survey research; the chapter on reliability and validity in Rosenthal and Rosnow (1991); Chapter 16 in Agresti and Finlay (1997); the monograph by Thurstone (1932); the previously-cited text by Traub (1994); the articles by Jackson (1939), Guttman (1953b), Cureton (1958), Lord (1959c), Novick (1966), Maxwell (1968), Traub (1997), van Belle and Arnold (2000); the clever "shoe size" note by Rogosa (2002); and the comparison of various reliability coefficients by Charter (2003).

A recent paper by Hershberger, Fisher, et al. (2005) addresses the problem of what happens to test-retest reliability if the error scores are correlated. They use as an example self-reported age of first drug use.

I am very comfortable with classical reliability theory, and so are many others, but there are a number of very vocal critics who are not. I already mentioned Loevinger (1957), Ross and Lumsden (1968), Lumsden (1976), and Cliff (1979); but there is also Guttman's (1953a) review of Gulliksen's book [with Gulliksen's (1953) rejoinder] and the article by Bock and Wood (1971), which is another critical review of classical reliability theory. Then there are those who propose generalizability theory (e.g., Cronbach, Gleser, et al., 1972; Shavelson & Webb, 1991; Brennan, 2001), item response theory (e.g., Hambleton, Swaminathan, & Rogers, 1991; van der Linden & Hambleton, 1997), or structural equation modeling (e.g., Bollen, 1989; Mueller, 1996) as alternative approaches to the theory of reliability for educational and psychological measurement. McDonald's (1999) textbook on test theory is a good source for comparisons of all of those approaches. Some of the principal differences between the concepts of classical reliability theory and those of generalizability theory, item response theory, and structural equation modeling will be explained briefly in Chapter 13 of this book.

CHAPTER 4: Attenuation

"Attenuation" is a fancy term for the reduction of the magnitude of a result due to the unreliability of measuring instruments. In this chapter I would like to talk about the effect of attenuation on various statistics (especially the correlation between two variables) and how one might correct for it.

What happens, and why

The effect is easy to state. If the axioms and theorems of classical reliability theory are satisfied, attenuation attributable to any amount of unreliability associated with measuring instruments lowers the correlation between any two variables such as height and weight, age and pulse rate, intelligence and achievement, or whatever. This can be shown by comparing the correlation between Obtained X and Obtained Y with the correlation between their true counterparts, i.e., r_{XY} vs. r_{TxTy}, as follows:

$r_{XY} = (\sum XY - NM_XM_Y) / NS_XS_Y$ [my favorite formulation for the Pearson correlation]

$= \{\sum(Tx + Ex)(Ty + Ey) - NM_{Tx}M_{Ty}\}/N \{(S_{Tx}/\sqrt{r_{XX}})(S_{Ty}/\sqrt{r_{YY}})\}$ [using Axiom #1 regarding the connection between obtained scores, true scores, and error scores; appealing to Theorem #1 regarding the equality of the mean obtained score and the mean true score, and solving for S_X and S_Y in the definitional formulas for the reliability coefficient for each of X and Y]

$= (\sqrt{r_{XX}}) (\sqrt{r_{YY}}) \{(\sum TxTy - NM_{Tx}M_{Ty})\} / NS_{Tx}S_{Ty}$ [re-arranging, and because when distributing the summation across the four terms in the product of Tx + Ex and Ty + Ey all of the terms drop out (are equal to 0) except for $\sum TxTy$]
$= (\sqrt{r_{XX}})(\sqrt{r_{YY}})\, r_{Tx,Ty}$ [using once again the now-familiar--I hope--formulation for the correlation between two variables, this time Tx and Ty]

But r_{XX} and r_{YY} are both between 0 and 1, and so are their square roots. Therefore, r_{XY} is always less than or equal to r_{TxTy}, i.e., the correlation that you actually get between two variables is a lower bound for the true correlation that you would have gotten if you had the true scores rather than the obtained scores. The effect is conservative. If there is any unreliability associated with either instrument, the user of the instruments must make weaker claims about the relationship between measurements obtained with those instruments. That is as it should be. It would be scientifically anomalous if you were able to find stronger relationships with unreliable instruments than with reliable ones. [Carroll (1997) does provide a non-intuitive example of a situation of “reverse-attenuation” for which the correlation between obtained scores is actually greater than the correlation between the corresponding true scores.]

The "correction"

The above derivation also leads to a "correction for attenuation". By solving for r_{TxTy} in the last step of the derivation we find that

$r_{TxTy} = r_{XY}/(\sqrt{r_{XX}})(\sqrt{r_{YY}})$ [if neither r_{XX} nor r_{YY} is equal to zero]..

So if you would like to estimate what the correlation might have been between True X and True Y, you would estimate the reliability coefficients of X and of Y (by knowing the correlation between the measurements on parallel forms of each), and divide the obtained correlation r_{XY} by the product of the square roots of the reliability coefficients.

Let's take an example (a hypothetical example with numbers that come out nicely). Suppose that for a group of 100 people you have obtained a correlation of .54 between their heights, X, and their weights,Y, when using a particular healthcare facility's scale, call it "Form A". Suppose further that you have also measured their heights and their weights using a second, "parallel", scale, call it "Form B", and the correlation between the Form A and Form B heights is .81 and the correlation between the Form A and Form B weights is .64. Then the "corrected for attenuation" estimate of the correlation between true height and true weight for those 100 people measured with those scales is equal to .54/ ($\sqrt{}$.81) ($\sqrt{}$.64) = .54/ (.90) (.80) = .54/.72 = .75. The obtained correlation that you did get (.54) is considerably less than the correlation that you should have gotten (.75) if you had the true heights and the true weights, because the latter has been "attenuated" by the less-than-perfect reliabilities of the measuring instruments. (There are actually four correlations between the heights and the weights involved here, viz., the correlation between the Form A heights and the Form A weights, the correlation between the Form A heights and the Form B weights, the correlation between the Form A weights and the Form B heights, and the correlation between the Form B heights and the Form B weights, but to strain this hypothetical example beyond credulity, let us assume that all of them are equal to .54.)

What can go wrong?

This is fine in theory, but in the real world the situation can get complicated and strange things can happen. First of all, we may only have one height&weight scale, not two. No problem, you say? Just measure everybody's height and weight twice with that same scale (the so-called test-retest or measure-remeasure approach), because the numbers won't know whether the "Form B" heights and weights were obtained with one scale or with interchangeable scales? Well, maybe, but remember what we need to say about errors of measurement. No matter how big or small they are, the errors associated with the second measurements are either assumed to be or can be shown to be uncorrelated with the errors associated with the first measurements, and the true

score is assumed to be the same for both forms. Both are less likely to be the case if only one "form" of an instrument is used (see, for example, Kelley, 1923; 1942).

There's more. For practical reasons you may be only able to measure the people's heights and weights once and must rely on other studies carried out on other people for estimates of the reliability coefficients for the heights and for the weights yielded by healthcare scales in general. (Worse yet, you may have to rely on evidence for the reliability of height and weight measurements using other devices such as yardsticks and inexpensive bathroom scales!) Suppose that to be the case here, and you find that Miller (I just made that up) reported estimated reliabilities of .49 and .36 for her healthcare scale heights and weights, respectively. (I assure you that they're not nearly that bad, but I'm trying to make a point.) Substituting those values for r_{XX} and r_{YY} in the correction-for-attenuation formula you get .54 / ($\sqrt{}$.49) ($\sqrt{}$.36), which is equal to 1.28, i.e., a reliability greater than 1. (See Tam & Knapp, 1997 for a similar example.) That wouldn't make sense, because a correlation coefficient cannot be greater than 1. (I hesitate to add that it is possible, although extremely unlikely, to get a NEGATIVE correlation between, say, the Form A weights and the Form B weights--if the axioms of reliability theory are not satisfied--rendering the situation equally anomalous, because one cannot find the square root of a negative value without getting involved with complex, i.e. imaginary, numbers.)

There's even more yet. If you have a matrix of estimated correlations between true scores for pairs of variables, and you try to use those correlations rather than the obtained correlations in a subsequent statistical analysis (a multiple regression analysis or a factor analysis, for example), it can happen that those correlations are not compatible with one another (the mathematical statisticians call such a matrix "non-Gramian"). That is, there may not exist any set of real numbers for which those correlations would be possible, in which case the desired analysis could "blow up" and not be capable of being carried out. (Those of you who are familiar with the "pairwise deletion" approach to missing-data problems should be aware that the same thing can happen there.)

<u>How many ways are there to get a particular correlation between two variables?</u>

Going back to the equation $r_{XY} = (\sqrt{r_{XX}}) (\sqrt{r_{YY}}) r_{TxTy}$, if the obtained correlation between two variables is equal to zero (unlikely, but possible), the usual conclusion is that there is no (linear) relationship between those variables. That may be, i.e., the correlation between true scores for the variables could be equal to zero, but it could also be the case that the true correlation is non-zero and either or both of the reliability coefficients for X and for Y are equal to zero, rendering that "triple product" equal to zero. At the opposite extreme, if the obtained correlation is equal to one (also unlikely, but also possible) all three of r_{XX}, r_{YY}, and r_{TxTy} must be equal to one, i.e, perfect reliability for both X and Y and a perfect linear relationship between their true scores. For any obtained

correlation between 0 and 1 there is an infinite number of combinations of r_{XX}, r_{YY}, and r_{TxTy} that could have produced that correlation. For example, if the obtained correlation is .25, it could be that the reliability coefficients for X and Y are both 1 and the true correlation is .25; the reliability coefficients are both .25 and the true correlation is 1; the reliability coefficient for X is .25, the reliability coefficient for Y is 1, and the true correlation is .5; or whatever. Do you see why a knowledge of the reliability of measuring instruments is very important? (Fleiss, 1986 felt it was so important that he devoted the very first chapter of his book, The design of clinical experiments, to "Reliability of measurement".)

The effect of attenuation on other statistics

Pearson correlations aren't the only statistics that can be attenuated by unreliability. Those who are familiar with the so-called "general linear model" know that there is a connection between correlations and differences between means. More specifically, if there is a big correlation between two variables X and Y then the difference between the mean Y for one level of X and the mean Y for another level of X is also big. So if we are interested in, say, the relationship between sex and height, we are more likely to concentrate on the degree of overlap between the frequency distribution of height for males and the frequency distribution of height for females than on the correlation between sex and height. It should come as no surprise to you, therefore, that any unreliability associated with the measurement of height will increase the degree of overlap between the two sexes (decrease the discriminability). That is, the amount of overlap that we actually get is greater than the true overlap.

Interestingly, although the Pearson product-moment correlation coefficient between an independent variable X and a dependent variable Y is attenuated by measurement error, the covariance between X and Y is not, and the regression coefficient (slope) in the regression of Y on X is attenuated by errors in X but not in Y. (See Bohrnstedt, 1983 for proofs of those assertions.)

Additional reading

The matter of attenuation was first brought to the attention of measurement specialists by Spearman (1904, 1907, 1910) and by Brown (1910, 1913). And there are a number of interesting discussions of attenuation in various research contexts (e.g., Thouless, 1939; Johnson, 1944, 1950; Murdaugh, 1981; Lee, Miller, & Graham, 1982; Bobko, 1983; Oumlil & Balloun, 1986; Mendoza & Mumford, 1987; Muchinsky, 1996; Schmidt & Hunter, 1996; Schmitt, 1996; Rogers, Schmitt, & Mullins, 2002) and its effects on specific statistics (e.g., Cochran, 1968; Fleiss & Shrout, 1977; Fuller & Hidiroglou, 1978; Winne & Belfry, 1982; Bohrnstedt, 1983--see above; Charles, 2005; Ree & Carretta, 2006; and Raju, Lezotte, & Fearing, 2006). Some educational and/or psychological measurement texts also have chapters or sections within chapters on attenuation.

CHAPTER 5: The interpretation of individual measurements

Although we started out by concentrating on the reliability of measurements taken on a single object (recall the situation regarding the measurement of a child's temperature using a less-than-perfect thermometer), the primary emphasis has been on the reliability of a measuring instrument for a GROUP of objects. Theorems #1, 2, and 3 were concerned with the mean obtained score vs. the mean true score ACROSS objects, the connection between obtained variance, true variance, and error variance ACROSS objects, and the estimation of the reliability coefficient ACROSS objects. In this chapter I would like to go back to the problem of determining the reliability of an individual obtained score and how one should interpret an obtained score that is "contaminated" by measurement error.

Back to our hypothetical example, and a little more theory

Consider the set of hypothetical obtained heights, true heights, and errors in Chapter 3, especially the first obtained measurement for Mary Smith: X = 60 inches. For those data we (with God's help) found that her true height for both of those interchangeable yardsticks was 64 inches. On that first measurement occasion we made an error of -4 inches. Will we always be off by four inches on the low side when using those yardsticks? Of course not. The measurement of Mary's height with the other yardstick produced an obtained score which just happened to be 64 inches, an error of 0 inches (we were lucky that time!). Additional obtained measurements would probably be numbers such as 66, 63, 68, etc., i.e., we would expect to get a distribution of obtained heights around Mary's true height of 64 inches. But what kind of distribution would that be, and how much would those obtained heights vary from one another? That requires another assumption.

Axiom #2 (a reasonable, albeit controversial, assumption): The obtained scores on parallel forms of an instrument for an individual object are normally distributed with the mean of those obtained scores equal to the individual object's true score and with the standard deviation of those obtained scores equal to the standard error of measurement for a group of objects, which is assumed to be the same regardless of the object's true score. (This also says that the corresponding error scores for an individual object are normally distributed around zero, since E = X - T, and T is a constant for that object.)

It is the first and the last parts of Axiom #2 that are controversial (the middle part is merely a re-statement of the definition of a true score). Why normally distributed? Why not, reply the measurement theorists. Shouldn't smaller errors be more frequent than larger errors, and large errors very unlikely? That's what happens in a normal distribution. But it is the assumption of the constancy of the standard error of measurement that most people find troubling. Surely, they say,

the standard error of measurement should be smaller for people who have very high true scores or very low true scores on an achievement test, for example, since there is a "ceiling effect" or a "floor effect", respectively, for their obtained scores; whereas the standard error of measurement should be larger for those people in the middle of the true score range (i.e., it should be “conditional” on true score). The resolution of that controversy ultimately comes down to various mathematical and empirical arguments regarding what happens in various portions of the obtained score range and estimating the standard error of measurement for each of those portions. Mollenkopf (1949--proof summarized on pp. 115-124 in Gulliksen, 1950) showed that the standard error of measurement is constant throughout the test score range if and only if the distribution of obtained test scores is symmetric and mesokurtic. (See also Lord, 1984; Woodruff, 1990; and Raju, Price, & Oshima, 2005. In their paper, Raju, Price, and Oshima extend the concept of conditional standard error of measurement to conditional reliability.)

For the time being we will adhere to all of the provisions of Axiom #2 and see what the implications are for interpreting individual obtained scores. In the measurement literature there are three approaches to the problem, all of which have counterparts in the basic concepts of inferential statistics.

How to interpret an individual measurement

Point estimation

The first approach is the point estimation of a true score. If we had to give one number that is in some sense our "best" estimate of an object's true measurement on a particular instrument, what would we say? (Her)his obtained score, X? No, that would be too liberal, because it would imply that we had a perfectly reliable measuring instrument. How about the mean obtained score (for a group of objects), M_X? No, that would be too conservative, because the mean obtained score is equal to the mean true score and would imply that our instrument were perfectly unreliable, i.e., it could not differentiate one true score for another. What is usually suggested is to report a "regressed score" (regressed toward the mean) determined as follows:

Estimated true score = $M_X + r_{XX} (X - M_X)$, where the reliability coefficient r_{XX} is estimated by r_{AB}, the correlation between parallel forms of the instrument (Kelley, 1927). [Wainer (2000) referred to something he called “Kelley’s Paradox” regarding what to use for M_X, i.e., “whose mean?”. Suen (1990) discussed the standard score version of this formula, and Payne (1989) provided an application of that formula to clinical psychology.]

For the hypothetical height data in Chapter 3, Mary Smith's true height would be estimated from her First X to be 70 + .50 (60 - 70) = 70 + .50 (-10) = 70 - 5 = 65. Using this formula, all "measurees" whose obtained scores are above the mean get estimated true scores that are less than their obtained scores and all

"measurees" whose obtained scores are below the mean (e.g., Mary) get estimated true scores that are greater than their obtained scores. (If the instrument were perfectly reliable, i.e., $r_{XX} = 1$, the estimated T for each person would be (her)his X; and if the instrument were perfectly unreliable, i.e., $r_{XX} = 0$, the estimated T for each person would be M_X, but that's not going to happen with real data.) The variance of those estimated true scores will therefore be less than the variance of the obtained scores, which is as it should be since the variance of the obtained scores is inflated by measurement error. That is not to say that the estimated true score for each measuree is "correct" or even in the appropriate relative position in the distribution of estimated true scores. All this does is provide us with best estimates of true scores "on the average". (Harris, 1973 provided a brief discussion of error defined as the discrepancy between estimated true score and "actual" true score, and contrasted that with both the traditional error of measurement, X - T, and with the error associated with the prediction of an obtained score on one form of an instrument from an obtained score on a parallel form.)

For the real data in Chapter 3 obtained with the Wright Peak Flow Meter, Subject 1's true volume would be estimated from (her)his First X to be 447.9 + .983 (494 - 447.9) = about 493 liters/minute. (The 447.9 "splits the difference" between the First X mean and the Second X mean for the entire group of 17 people.)

Interval estimation

The second approach should be familiar to the reader who has studied confidence interval estimation in basic statistics. Recall that a particular statistic, call it A, such as a mean, a standard deviation, or a Pearson r, is often interpreted as A plus or minus some "margin of error". For individual measurements we can do the same thing. Here we take an object's obtained score as the statistic and some multiple of the standard error of measurement as the margin of error. Going back to the hypothetical data for the heights of those seven people, we see that Mary Smith's first obtained height is 60 inches and the standard error of measurement, S_E , is 4 inches. (Note that the standard error of measurement is "scale-bound" in the units of measurement for the variable in which we are interested, unlike the reliability coefficient, which is "scale-free".) We therefore report that Mary is 60 ± 4 inches tall or, if we want to be more conservative, that she is 60 ± 2(4) = 60 ± 8 inches tall, giving ourselves an even greater margin of error; or whatever.

The ± 4 and the ± 2(4) come from the normal distribution (the 2 is actually 1.96), corresponding to 68% confidence and 95% confidence, respectively. But extreme caution must be observed here (unless you're a Bayesian). Axiom #2 claims that an individual's obtained scores are normally distributed around her(his) true score. When we report a confidence interval for Mary Smith's true height as 60 ± 8 = 52 to 68 we must be very careful to say something like "52 inches and 68 inches are reasonable limits for Mary's true height" and NOT "the

probability is .95 that Mary's true height is between 52 inches and 68 inches". True scores, though unknown (in the real world, that is), are fixed and are not distributed around obtained scores; it is the obtained scores that are distributed around the true scores.

Subject 1 in the Wright Peak Flow Meter data would have a 95% confidence interval for (her)his true score of 464 to 524 with respect to (her)his obtained score of 494.

Hypothesis testing

The third approach should be the most familiar of all. Instead of giving one number that is the best estimate of a true score, or two numbers that are reasonable limits for a true score, you test a hypothesis about the true score, and on the basis of the test you either reject or do not reject that hypothesized value. Here's how it goes for true scores (again using our example of Mary Smith's first obtained height):

Null hypothesis: $T = 66$ (say) [T is the unknown parameter]
Alternative hypothesis: $T \neq 66$ [a non-directional, "two-sided", alternative]

$X = 60$ [X is the obtained statistic.]

$S_E = 4$ [obtained from $S_E = S_X \sqrt{(1 - r_{AB})}$]

Test statistic = $(X - T) / S_E = (60 - 66) / 4 = (-6) / 4 = -1.50$

If that test statistic is assumed to be normally distributed (in accordance with Axiom #2), we cannot reject the hypothesis that T = 66, because a ratio of -1.50 is within "the acceptance region" for the standardized normal sampling distribution. In other words, the obtained score of 60 is "close enough" to a true score of 66 for this unreliable yardstick so that 66 is not "rejectable". This is NOT the same as claiming that T is equal to 66. We still don't know what it is (which is of course the frustrating aspect of inferential statistics in general).

For the peak flow data, if 400 were hypothesized to be Subject 1's true score, that hypothesis would be soundly rejected, because (494 - 400)/ 14.9 = 6.31 is well beyond the "critical ratio" of 1.96 for the .05 significance level.

Note that the interval estimation approach actually subsumes the hypothesis testing approach (as is most often, but not always, the case). A true score of 66 is within the 95% confidence interval of 52 to 68, and it is therefore not rejectable as a candidate for Mary's true height if you adopt .05 as the level of significance. Had it been outside the limits of that interval, the null hypothesis of T = 66 would have been rejected in favor of the alternative hypothesis that $T \neq 66$.

Compounded measurement error

In the physical sciences it is often the case that two or more less-than-perfectly-reliable measurements are made on the same object and then those measurements are combined mathematically to produce a third measurement, which is necessarily also less than perfectly reliable. Formulas have been developed by researchers affiliated with the National Institute of Standards and Technology (see, for example, Taylor & Kuyatt, 1993) for determining interval estimates of the true scores for such measurements as functions of the standard errors of measurement for the individual components. Topping (1975) has provided similar formulas and my friend John Pezzullo has an interactive webpage (on Propagation and Compounding of Errors) where you can input the obtained score and standard error of measurement for each of two measurements and receive as output the resulting expression and its standard error.

As illustrations of how this works, consider body surface area (BSA) and body mass index (BMI). One fomula for body surface area (DuBois & DuBois, 1916) is the constant .20247 times height (in meters) raised to the .725 power times weight (in kilograms) raised to the .425 power. Body mass index (the Quetelet index) is equal to weight (in kilograms) divided by the square of height (in meters). Suppose you would like to get 95% confidence intervals for true body surface area and true body mass index for our hypothetical friend, Mary Smith. You measure her height and get 60 inches; you measure her weight and get 120 pounds. Her obtained body surface area is 1.50 square meters and her obtained body mass index is 23.4 kilograms per square meter. Your height measuring instrument is said to have a standard error of measurement of 4 inches (that's awful--see Chapter 3) and your weight measuring instrument is said to have a standard error of measurement of 5 pounds (that's also awful); so the 95% confidence interval for Mary's true height is 60 ± 2(4) or from 52 inches to 68 inches, and the 95% confidence interval for Mary's true weight is 120 ± 2(5) or from 110 pounds to 130 pounds.

According to Taylor and Kuyatt, if Y (the quantity you're interested in) is equal to any constant A times the product of X_1 raised to the power a and X_2 raised to the power b, then you can determine the "uncertainty" (using their term for standard error of measurement) associated with Y by the following formula:

$$\text{Uncertainty of } Y = \left[a^2(SE_{X1} / |X_1|)^2 + b^2(SE_{X2} / |X_2|)^2\right]^{.5}$$

where |Y| is the absolute value of Y, SE_{X1} is the standard error of measurement for X_1 , $|X_1|$ is the absolute value of X_1 , SE_{X2} is the standard error of measurement for X_2 , and $|X_2|$ is the absolute value of X_2 , if both X_1 and X_2 are not equal to zero.

For body surface area, if height = X_1 and weight = X_2 , then A = .20247, a = .725, and b = .425. For body mass index, if again height = X_1 and weight = X_2 , then A = 1, a = 1, and b = -2. Substituting in the standard error (uncertainty) formula for Y and laying off two standard errors around the obtained BSA and the obtained BMI, we have

Body surface area: 1.50 ± 2 (.05) = 1.40 to 1.60

Body mass index: 23.5 ± 2 (3.3) = 16.9 to 30.1

Body surface area is often used as the basis for determining the appropriate dose of medication to be prescribed (BSA is multiplied by dose per square meter to get the desired dose), so you can see from this admittedly extreme example that reasonable limits for "the true required dose" can vary dramatically, with possible serious medical complications for a dose that may be either too large or too small.

Body mass index is often used for various recommended weight therapies, and since the lower limit of the 95% confidence interval for Mary's true BMI is in the "underweight" range and the upper limit is in the "obese" range, the extremely high standard errors of measurement for both height and weight had a very serious effect on BMI. (Thank goodness these are hypothetical data for very poor measuring instruments.)

Isn't reliability fascinating?

Additional reading

For more on the interpretation of individual obtained measurements and inferences regarding their true counterparts, see Gulliksen (1950), Lord (1959b, 1959d), Zimmerman and Williams (1966), Cronbach (1970), Knapp (1971), Dudek (1979), Jarjoura (1985), and Huynh (1986a). The article by Dudek is especially relevant. He argues that there are three different standards errors of measurement and that it is essential to use the right one at the right time.

CHAPTER 6: The reliability of difference scores

The determination of the reliability of an obtained score is a difficult task, as we have seen. But how about the determination of the reliability of the difference between two obtained scores? Scientists have always been interested in differences. I've already mentioned (in Chapter 4) the difference between the mean height of males and the mean height of females. Another example that immediately comes to mind is the difference between the percentage of subjects in an experimental (treatment) group who experience pain relief and the percentage of subjects in a control (placebo) group who experience pain relief in a randomized clinical trial for a new drug. And there are lots of others.

Types of difference scores

As far as scientific measurement is concerned, there are five types of difference scores whose reliability seems to be of greatest interest. They are:

(1) the difference between an obtained score for an object on a particular instrument at one point in time and another obtained score for that same object on that same instrument at approximately the same point in time;

(2) the difference between an object's obtained score on a particular instrument and another object's obtained score on that same instrument at approximately the same point in time;

(3) the difference between an object's obtained score on a particular instrument at one point in time and the object's obtained score on that same instrument at a subsequent point in time;

(4) the difference between an object's obtained score on a particular instrument and the object's obtained score on another instrument at approximately the same point in time;

(5) the difference between an object's obtained score assigned by one rater using a particular instrument and the object's obtained score assigned by another rater using that same instrument at approximately the same point in time; or the difference between an object's obtained score assigned by one rater using a particular instrument at one point in time and that same rater using that same instrument at a subsequent point in time.

The first of these is relevant to what we have called test-retest or measure-remeasure reliability, and it is concerned with questions such as "Is the difference between Mary's spelling test score and immediate retest score within acceptable bounds of measurement error?" and "Is the difference between Mary's two simultaneously obtained weights indicative of a problem with the scale?"

The second type of difference score could be called "an inter-object discriminability score" (I just made up that term), and it is concerned with questions such as "How much better did Mary perform on a spelling test than John did?" and "How much more does John weigh than Mary does?"

The third type is often called a "simple change score" or "simple gain score" (where a negative gain is a loss) and it is concerned with questions such as "How much did Mary gain in reading achievement from the beginning of the school year to the end of the school year?" and "How much weight did John lose from the beginning of the dieting experiment to the end of the experiment?" There are several varieties of change scores--see below.

The fourth focuses on the discrepancy between the measurements produced with two instruments that are alleged to measure the same thing or with two instruments that are alleged to measure different things, and it is concerned with questions such as "How much higher is Mary's temperature taken with an electronic thermometer than with a traditional mercury-in-glass thermometer?" and "How much better did John perform on a spelling test than he did on a vocabulary test?"

The fifth type of difference score is the inter-rater type of score (inter-judge and inter-observer are common synonyms), or intra-rater (intra-judge, intra-observer) type of score, and it is concerned with questions such as "When Mrs. Jones and Mr. Brown both graded Mary's essay examination, how much did those grades differ?" and "When Mrs. Jones and Mr. Brown both measured John's weight, how much did those weights differ?"; or with questions such as "When Mrs. Jones graded Mary's essay twice, how much did those grades differ?" and "When Mr. Brown measured John's weight twice, how much did those weights differ?"

As I mentioned above, there are several kinds of change scores that have been reported in the literature, e.g., modified gain scores ("How much actual gain was there out of how much possible gain?"), percent change scores ("How much change was there relative to the initial measurement?"), optimally-weighted change scores ("What is our best estimate of true change?"), and residual change scores ("How much change was there over and above what was predictable?"). As you will see, it is the simple change score and the residual change score that have been the subjects of the most controversy.

<u>The general case</u>

The reliability of any difference score, Y - X, can be determined, or at least estimated, by using the following formula (see Stanley, 1967 for its derivation, but in different notation):

$$r_{Y-X,\,Y-X} = \frac{r_{AB}S_AS_B + r_{CD}S_CS_D - r_{AD}S_AS_D - r_{BC}S_BS_C}{\sqrt{(S_A^2 + S_C^2 - 2r_{AC}S_AS_C)}\,\sqrt{(S_B^2 + S_D^2 - 2r_{BD}S_BS_D)}}$$

where r_{AB} and r_{CD} are the Pearson correlations between obtained scores on Forms A and B of X, and between obtained scores on Forms C and D of Y, respectively (and therefore estimates of the reliability coefficients for X and for Y); the other r's are the Pearson correlations between the obtained scores for the forms indicated by their subscripts; and the S's are the standard deviations of the obtained scores for the forms also indicated by their subscripts.

This "monster" formula (as Stanley, 1967 calls it--since it has six correlations and four standard deviations to contend with!) can be simplified by making certain assumptions or having certain evidence regarding its various components, depending upon what type of difference is of interest. Let us consider the five basic types of difference scores in order, concentrating on how that formula might be simplified, and what some of the implications are for interpreting a particular difference score.

Measure-remeasure differences

The first of the above types is of primary interest in the physical sciences. If boxes of various sizes are "double-measured" for length, width, and depth, for example, and there are serious within-dimension within-box discrepancies in those measurements, there is a problem with the measuring instrument. It is not of primary interest in the social sciences, however, because immediate retesting might seem rather silly. Referring again to the spelling test illustration, it would be at least unusual to give Mary a spelling test, pick up her paper, and then give her the same test again! In any event, we need not have any special formula for the reliability of such difference scores, since their reliability is r_{XX} itself, which can be estimated by r_{AB} where A is the first obtained score and B is the essentially simultaneous second obtained score, and their parallelism is accordingly assumed. (There might be a problem of correlated errors, however; see Chapter 3.) But the standard error of the difference is greater than the standard error of measurement by a factor of $\sqrt{2}$.

As an illustration of this type of difference score, consider our hypothetical height data (see Chapter 3 and Chapter 5) and Mary's obtained heights of 60 inches and 64 inches. For this instrument with a reliability coefficient of .50, the standard error of measurement (for a single obtained score) was found to be 4 inches. Therefore the standard error of the difference between two obtained scores for the same object is $\sqrt{2}\,(4) = 1.414(4) = 5.656$. The discrepancy of 4 inches between her two obtained height measurements is well within the "margin of error" for that instrument.

Between-object differences

The second type of difference score is of general concern in all sciences. The emphasis there is on deciding such things as whether Mary's obtained score on a spelling test is enough higher than John's obtained score on the same test for us to conclude that there is a difference between their corresponding true scores on that test. In those situations the general formula for the reliability of a difference score also reduces to r_{XX} (i.e., r_{AB}). This is not intuitively obvious, so it must be proven, but rather than trying to wrestle with various assumptions about the six correlations and the four standard deviations in the general formula, it can be demonstrated in the following way:

The variance of the difference between the true score for object i and the true score for object j on the same instrument, i.e., the variance of $T_i - T_j$ for all i and j (i,j =1,2,…,N) is (1/N) $\{ \sum[(T_i - T_j) - M_{Ti - Tj}]^2\}$, from the definition of a variance. Expanding, distributing the summation sign, and multiplying each term by 1/N, we get (1/N) $\sum(T_i - T_j)^2$ - (2/N) $M_{Ti - Tj} \sum(T_i - T_j)$ + (N/N) $M_{Ti - Tj}^2$. But the last two expressions are both equal to 0, since $M_{Ti - Tj} = 0$. (All objects are listed N times for i and N times for j, and the mean of all of those differences is equal to the difference between the mean for i and the mean for j, and that difference must be 0.) It can also be shown that the variance of ANY variable A can be calculated using the formula (1/2N^2) $\sum (A_i - A_j)^2$, for all i and j. (Another formula for a variance? Yes; trust me.) Letting T = A and simplifying, we have that the variance of $T_i - T_j$ is equal to $2NS_T^2$. It can be similarly shown that the variance of $X_i - X_j$ is equal to $2NS_X^2$. Therefore the reliability of this particular type of difference score is $2NS_T^2/2NS_X^2$, and the 2N in the numerator and the 2N in the denominator "kill each other", leaving S_T^2/S_X^2 or r_{XX}, i.e., r_{AB}.

This means that the "over-all" reliability of the difference between two obtained scores for a particular instrument is the same as the "over-all" reliability of an individual obtained score, but the standard error of that difference is also larger than the standard error of measurement by a factor of $\sqrt{2}$. (The error variance doubles and so does the obtained variance, so the reliability coefficient remains the same.) To illustrate this, consider again our hypothetical height data. If our primary interest were in estimating Mary's true height from her first obtained height, the 95% confidence interval would be 60 ± 2(4), i.e., from 52 to 68. But if our primary interest were in estimating the difference between Mary's true height and Carol's true height, the 95% confidence interval would be (60 -64) ± 2($\sqrt{2}$)(4) or -4 ± 11.312, i.e., from -15.312 to 7.312, a much larger "margin of error". Although there is an obtained difference in their heights of 4 inches (with Mary shorter than Carol), the true difference in their heights could reasonably range from Mary being 15.312 inches shorter than Carol to Mary being 7.312 inches taller than Carol. Pretty lousy measuring instrument, isn't it? (But we already established that.)

Change scores

Simple change

For the reliability of a difference score that is a simple change score, you might be relatively safe in assuming that the two within-X variances are equal to one another (since the A and B forms are parallel), the two within-Y variances are equal to one another (since the C and D forms are also parallel), and the four "cross-correlations" between X and Y(r_{AC}, r_{AD}, r_{BC}, and r_{BD}) are all equal. You wouldn't be nearly as safe in assuming that the reliability coefficient for X and the reliability coefficient for Y are equal--there might be better or worse consistency at Time 2 than there is at Time 1; or that the X-variances and the Y-variances are equal to one another--the change in level (e.g., with most subjects getting higher scores) might also be accompanied by a change in dispersion (e.g., with some subjects making small gains and other subjects making large gains). If you do make those assumptions, the formula reduces to

$$r_{Y-X,\ Y-X} = \frac{r - r_{XY}}{1 - r_{XY}}$$

where r is the reliability coefficient for X and for Y, and r_{XY} is the correlation between obtained scores for X and obtained scores for Y. Investigations of that formula and its possible implications have produced one of the most heated controversies in scientific measurement.

Controversy regarding the measurement of simple change

The most common conclusion drawn from an investigation of that formula is that the reliability of a simple change score is disappointingly low, for desirable and hopefully-typical values of r and r_{XY}. If, for example, r is .80 and r_{XY} is .50, then the reliability of Y - X is (.80-.50)/ (1-.50) = .60; i.e., the reliability of the difference is less than the reliability of either X or Y taken separately. But is an r of .80 desirable and typical? Yes, we want the reliability of X and the reliability of Y to both be high, and many instruments in common use have reliabilities of .80 or higher. How about an r_{XY} of .50? Yes, if X and Y are not reasonably highly correlated (and values in the general magnitude of .50 are commonly found for Time 1 and Time 2 measurements taken on the same objects) we should not be subtracting X from Y; it would be tantamount to subtracting apples from oranges. (But see Bereiter, 1963 and Willett, 1988-1989 for the opposing viewpoint regarding the necessity for r_{XY} to be large. Bereiter also provided a formula for estimating the reliability of total-score change as a function of the differences between obtained scores for corresponding items on X and Y--he called them "change items".)

The foregoing argument concerning the unreliability of change scores has been so impressive to some people that it has played a major role in their advocacy of not only doing away with the simple change score itself but with the entire concept of change (see, for example, Cronbach & Furby, 1970; O'Connor, 1972; Linn & Slinde, 1977; and my articles--Knapp, 1980, 1984a). Cronbach and Furby were also influenced by the argument that X and Y might measure different psychological processes at the two different time points (e.g., aptitude at Time 1 and achievement at Time 2) and by the argument that we often don't need to talk about change, even in a controlled experiment employing a pretest and a posttest, given what we know about the analysis of covariance and its applications to "over-and-above the effect of the pretest" situations. (See Maris, 1998 for a recent discussion of the use of simple change scores vs. the analysis of covariance.)

Critics of the argument that simple change scores are necessarily unreliable (e.g., Maxwell & Howard, 1981; Zimmerman & Williams, 1982; Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1983; Williams & Zimmerman, 1984; Rogosa, 1988; Willett, 1988-1989; Zimmerman, 1994; Rogosa, 1995; Zumbo, 1999) have presented counter-arguments that have taken a variety of forms. In their critique of my 1980 article, Williams and Zimmerman (1984) disagreed with some of the claims that I made and took exception to my hypothetical numerical example (essentially the same example I used in Chapter 3 of this book, but with linearly-transformed numbers and a different substantive context). They provided a different hypothetical numerical example that they regarded as superior, but in my response (Knapp, 1984a) I begged to differ. They also argued elsewhere (e.g., Williams & Zimmerman, 1977) that the assumptions underlying classical reliability theory are not reasonable for the measurement of simple change, since the error scores at Time 1 are likely to be correlated with the error scores at Time 2 when you're measuring change from Time 1 to Time 2.

Rogosa et al. (1982), Rogosa and Willett (1983), Rogosa (1988), Willett (1988-1989), and Rogosa (1995) urged the adoption of a multi-wave, longitudinal "growth curve" approach to the measurement of change, rather than a simple two-wave (Time 2 minus Time 1) approach. Willett's article, which is an excellent summary of the measurement of change in general, contains an informative artificial example of data collected on the same subjects at four timepoints. It has been shown, by Heise (1969) and by Wiley and Wiley (1970, 1974), that if you have test-retest data for four points in time there is a method for distinguishing between the stability of the instrument (reliability) and the stability of the unknown true scores (true growth), and for testing the assumptions all in one full swoop. (With data for only two points in time, instrument stability and true score stability are confounded with one another.) Two co-authors and I described that method in an article in Nursing Research (Knapp, Kimble, & Dunbar, 1998) and included the following real-data example for four repeated administrations of the monopolar version of the Profile of Mood States to 46

cardiac dysrhythmia patients (at entry to the study, one month later, three months later, and six months later)--the subscripts refer to the time points:

r_{12} = .686, r_{13} = .602, r_{14} = .542, r_{23} = .829, r_{24} = .770, and r_{34} = .893

Applying Heise's method to these correlations, we found that our "best" estimate of the reliability of the instrument is .945 and our "best" estimates of the stabilities of the underlying true scores for the first three time points are s_{12} = .726, s_{13} = .637, and s_{23} = .878; i.e., good reliability in general and greatest stability between times 2 and 3.

Zumbo (1999) joined Williams and Zimmerman's defense of simple change scores, essentially agreed with Rogosa, Willett, and others about the superiority of the multi-wave approach (and advocated the incorporation of structural equation modeling into the analysis), but also provided handy guidelines for researchers who for practical reasons are restricted to just two waves.

Modified change

Other people have suggested alterations to the simple change score that may make better sense and/or may have better reliabilities. The easiest (to think about, anyhow) of these is the so-called "modified gain score" (Y - X)/ (K - X), where K is the maximum possible obtained score for a given instrument. Advocates of the modified gain score are concerned not so much about the unreliability of simple change scores as about the "ceiling effect" for many educational and psychological tests. A two-point gain from 50 to 52 on a 100-item test, for example, should not be treated the same as a two-point gain from 96 to 98 on that same test, in their opinion. (The modified gain scores would be 2/50 = .04 and 2/4 = .50, respectively.) Unfortunately, a formal statistical comparison of the relative reliabilities of simple change scores and modified change scores has never been carried out, as far as I can determine, but I think it would be a mess, for two reasons: (1) derivations for statistics that are products or quotients of variables (the latter being the case here) have always been more difficult than similar derivations for sums or differences; and (2) the K - X in the denominator causes serious problems whenever K = X (i.e., when the obtained score for an object at Time 1 is equal to the maximum possible score), since for that object the modified gain score would be undefined (you can't divide by 0). In any event, I can't give you any simplified formula for the reliability of a modified gain score.

Percent change

Another variation is percent change (see VanMeter, 1974), 100 (Y - X)/X. VanMeter gives the example of change in compensation for state legislators over a five-year period in the 1960s, where the raw gain ($7200) for legislators in Missouri (from $11550 to $18750) was slightly larger than the raw gain ($6975)

for legislators in Tennessee (from $1225 to $8000) but the percent gain for the former was 62.3% compared to 620% for the latter. But like the modified gain score, a formal statistical comparison has not been made between the reliability of a simple change score and the reliability of a percent change score. I'm afraid it would also be a mess, however, for essentially the same reasons I just gave for modified gain scores: (1) the measure is a quotient of two variables; and (2) you still have the problem of division by 0, if the initial measurement is 0.

Weighted change

A much more complicated alteration is the "optimally-weighted change score". Optimally-weighted scores have a long history, originating (I think) with the work of Mosier (1943) and carried through by Gulliksen (1950), Lord (1956;1958), McNemar (1958), Stanley (1971), and many others. The problem is one of determining how to weight obtained Y and how to weight obtained X in a linear composite of the two in an expression of the form aX + bY + c that provides "the best" estimate of the difference between true Y and true X, i.e., true change. The optimal weights and constant term turn out to be

$$a = \{1/(1 - r_{XY}^2)\} \{(S_Y/S_X)\, r_{XY}\, (1 - r_{YY}) - r_{XX} + r_{XY}^2\}$$

$$b = \{1/(1 - r_{XY}^2)\} \{(S_X/S_Y)\, r_{XY}\, (r_{XX} - 1) + r_{YY} - r_{XY}^2\}$$

$$c = M_Y - M_X - aM_X - bM_Y$$

[I told you it was complicated!]

and the reliability of the difference is

$$r_{Y-X,Y-X} = \frac{r_{XX} S_X^2 + r_{YY} S_Y^2 - 2r_{XY} S_X S_Y}{S_X^2 + S_Y^2 - 2r_{XY} S_X S_Y}$$

where r_{XX} and r_{YY} are the reliabilities of X and Y, respectively (and are to be estimated by r_{AB} and r_{CD}, again respectively); r_{XY} is the correlation between X and Y (to be estimated by r_{AC} and/or r_{BD}); S_X is the standard deviation of X (to be estimated by S_A and/or S_B) and S_Y is the standard deviation of Y (to be estimated by S_C and/or S_D); and M_X and M_Y are the corresponding means (to be estimated by M_A and/or M_B and M_C and/or M_D, respectively).

Residual change

Another complicated alteration to the simple change score is the "residual change score". Here the problem is the determination of an estimate of true change that is over and above what one could obtain by regressing obtained Y on obtained X ("predicting" obtained Y from obtained X). And you can get

different answers for the reliability of a residual change score depending upon whether you want residual change to be uncorrelated with initial X or with initial T. That matter has produced another controversy (see, for example, DuBois, 1957; Manning & DuBois, 1962; Bechtoldt, 1963; Tucker, Damarin, & Messick, 1966; Traub, 1967, 1968; Glass, 1968; Bond, 1979). There are accordingly two expressions for the reliability of a residual change score. The first (favored by Manning and DuBois), is

$$r_{Y-(a+bX),Y-(a+bX)} = \frac{r_{YY} - r_{XY}^2 (2 - r_{XX})}{1 - r_{XY}^2}$$

where Y-(a+bX) is the notation for a residual change score (a is the Y intercept and b is the regression coefficient for predicting Y from X), r_{YY} is the reliability of the obtained scores at Time 2, r_{XX} is the reliability of the obtained scores at Time 1, and r_{XY} is the correlation between obtained scores at Time 1 and obtained scores at Time 2. The assumption here is that residual change is uncorrelated with initial X.

The second formulation (favored by Tucker, Damarin, and Messick), is

$$r_{Y-(a+bX),Y-(a+bX)} = \frac{r_{XX} (r_{XX} r_{YY} - r_{XY}^2)}{r_{XX}^2 - 2r_{XY}^2 r_{XX} + r_{XY}^2}$$

The quantities in this expression are the same as those in the previous expression, but it is assumed that residual change is uncorrelated with initial T, not with initial X.

Other difference scores that are not change scores

But enough (for now) about change scores of various kinds. Let us move on to discuss the other popular types of difference scores.

Inter-instrument differences

When Y - X is the difference between an object's obtained scores for two instruments that are alleged to measure the same thing, by transforming all of the "raw" obtained scores into standard scores (a perfectably acceptable thing to do, since Y and X may not have the same metric) and asssuming that $r_{BD} = r_{AC}$ and $r_{BC} = r_{AD}$, we have

$$r_{Y-X,Y-X} = \frac{1/2 (r_{AB} + r_{CD}) - r_{AD}}{1 - r_{AC}}$$

The situation is essentially the same statistically (but not substantively) when Y - X is the difference between an object's obtained scores for two instruments that are alleged to measure different things. The assumptions just made are likely to be equally justified, and the simplified formula for estimating the reliability of that type of score is the same.

Inter- and intra-rater differences

If all measuring instruments were perfectly "objective", requiring no human intervention whatsoever, science (and life) would be much simpler. But there are many situations, especially in educational and psychological measurement, where personal judgments must be made in the process of obtaining the measurements. The most common of such situations are those in which human beings evaluate other human beings. In Chapter 1 I referred to the example of teachers who grade essays that have been written by their pupils. The teachers themselves are the "instruments"; their pupils are the "objects". The obtained scores are those produced by the "raters" (teachers), and those obtained scores may or may not be good approximations to the true scores that the "ratees" deserve to get. And as indicated in that chapter and in the list of types of difference scores above, the second rating might be provided by a different teacher at approximately the same time or by the same teacher at a subsequent time. For the inter-rater case, any assumptions about the equality of various correlations and/or variances other than the within-rater standard deviations are hazardous, so the formula for the reliability of inter-rater differences remains a "monster".

Things are a bit nicer for the intra-rater case. Additional assumptions such as $r_{AC} = r_{BD}$ would appear to be warranted but the assumption that $r_{AB} = r_{CD}$ probably would not, since there could be "slippage" in reliability at Time 2 (due to fatigue, for example), even though the rater is the same.

Our flow meter example (revisited)

In order to further illustrate the estimation of the reliability of some of the above types of difference scores I'd like to close this chapter by returning to our real-world example concerning the measurement of expiratory flow rate (Bland & Altman, 1986) that was discussed in Chapter 3 and Chapter 5. Recall that we had four columns of data: (1) measurements taken on 17 people with the standard Wright peak flow meter; (2) repeated measurements taken on those people at approximately the same time with the same meter; (3) measurements taken on the same people with a "mini" version of that meter; and (4) repeated measurements taken on the same people with the "mini" meter. Those data lend themselves to the study of three of the types of difference scores (#1, #2, and #4 in my list of the five types of greatest interest), with the four columns of data corresponding to the variables A, B, C, and D of Stanley's (1967) formula.

Consider first the matter of the reliability of the difference between the double measurements taken on the same person with the standard meter, e.g., the 494 and the 490 for Subject 1. According to the argument given above, the "over-all" reliability is the same as the reliability of a single obtained measurement, i.e., .983 (see Chapter 3 for the summary statistics for those data), but the standard error of the difference is □2 times larger than the standard error of measurement, i.e., 1.414 (14.9) = 21.1. The difference of 4 liters/minute between Subject 1's obtained measurements is well within that "margin of error".

Next, consider the difference between the first measurements taken on different people with the standard meter, e.g., the 494 and the 395 for Subject 1 and Subject 2, respectively. Is that difference of 99 liters/minute big enough for us to claim that those two people have different true scores? According to the same argument just used, the "over-all" reliability of such differences is the same as the reliability of a single obtained measurement taken with that meter, i.e., .983, but the standard error of the difference is again √2 times larger than the standard error of measurement for a single obtained score, i.e., 21.1. "Laying off" twice that amount (for 95% confidence) around the obtained difference of 99, we get 99 ± 42.2 or 56.8 and 141.2 as "reasonable limits" for the difference between the true scores for Subject 1 and Subject 2. Since that interval does not include 0, we can rest comfortably assured that those two subjects do not have equal true scores. [Exercise for the reader: How about the difference between Subject 1 and Subject 3 at Time 1 for that meter?]

Now consider the difference between measurements taken on the same person with two different instruments, e.g., the 494 for Subject 1 for the standard meter and the 512 for Subject 1 for the mini meter. Is that difference of 18 liters/minute big enough for us to claim that Subject 1 has different true scores for those two instruments (which are actually alleged to measure the same thing)? Referring back to the appropriate formula for the reliability of a difference score such as this, and substituting the corresponding values for the flow meter data, we have

$$r_{Y\text{-}X,Y\text{-}X} = \frac{1/2\,(.983 + .967) - 1/2\,(.936 + .957)}{1 - 1/2\,(.943 + .952)}$$

= .547, which is somewhat discouraging (but those ARE different instruments), and the corresponding standard error of the difference is $S_{Y\text{-}X}\sqrt{(1 - r_{Y\text{-}X,Y\text{-}X})}$, which turns out to be (trust me) 36.4 (.673) = 24.5. The obtained difference of 18 is within that "margin of error", so we cannot conclude that Subject 1's true scores for the two instruments are different. In other words, both meters appear to be getting at the same thing, at least as far as Subject 1 is concerned.

[The simplified formula for the reliability of that type of difference score had only r_{AD} as the last term in the numerator and only r_{AC} as the last term in the

denominator, but for the flow meter data r_{BC} was not the same as r_{AD}, and r_{BD} was not the same as r_{AC} , so they had to be averaged. Likewise for the variances in the estimation of $S_{Y-X,Y-X}$. Do you follow all of that?]

Additional reading

There is a vast literature on the reliability of difference scores of various types. Among the sources that have not been cited previously in this chapter, I recommend two chapters in the book edited by Harris (1963) in addition to the already-cited chapter by Bereiter (one by Lord, and one by Webster & Bereiter); the articles by Bohrnstedt (1969), Zimmerman, Brotohusodo, and Williams (1981), Cattell (1982), Gardner and Neufeld (1987), Guyatt, Walter, and Norman (1987), Malgady and Colon-Malgady (1991), and Guyatt, Kirshner, and Jaeschke (1992); the articles by Coleman and by Siegel and Hodge in the book edited by Blalock and Blalock (1968); the articles by Cardinet and by Embretson in the book edited by Laveault, et al. (1994); the article by Edwards (2001); the article by Hertzog and Nesselroade (2003), and the guidelines by Hummel-Rossi and Weinberg (1975). Edwards discusses what he calls "myths" concerning difference scores. I leave it to you to decide whether they are or are not.

CHAPTER 7: The reliability of a single item

One thing that differentiates physical science instruments from social science instruments is that the former usually do not have "items". When you measure something like the width of a box, you don't have a bunch of items that, taken together, produce a width measurement. When you measure something like mathematical ability, however, you invariably have one or more test items that constitute an instrument for measuring such an ability.

This chapter and the following chapter will be of little or no interest to you if you are concerned primarily with physical measurement. But if you are concerned primarily with social measurement, they should be of considerable interest. In this chapter I want to discuss the case of a single-item instrument. In Chapter 8 I will turn to the case of multi-item instruments, i.e., instruments consisting of two or more items.

Single-item examples

What do I mean by a single item? The simplest example is "Who is the president of the United States?" Another example is "Do you agree with the 1973 Supreme Court decision concerning abortion? (a) Yes; (b) Undecided; (c) No". Items such as these are often "stand-alone" instruments, and we need to be able to estimate their reliability every bit as much as we need to estimate the reliability of thermometers and yardsticks. But the theoretical underpinnings for obtained scores, true scores, and error scores will be a bit different, as we shall see, principally because for single items we usually do not have the luxury of continuous or "continuous-enough" metrics that permit the usual operations of addition and subtraction that permeate classical reliability theory.

X, T, and E for single dichotomous items

Consider the "Who is the president of the United States?" item, with dichotomous scoring (right answer = 1; wrong answer = 0). The concept of obtained score, X, presents no problem. There are two possible obtained scores: 1 (if you get it right); and 0 (if you get it wrong). The concept of true score, T, is almost as straightforward. There are two possible true scores: 1 (if you "deserved" to get it right); and 0 (if you "deserved" to get it wrong). The concept of error score, E, gets a little tricky, however. If we assume that $X = T + E$ as we did for continuous variables, then whenever $X = T$ (when they're both equal to 1 or both equal to 0), E is equal to 0; whenever $X \neq T$ (when $X = 1$ and $T = 0$, or when $X = 0$ and $T = 1$), E is either 1 or -1. Therefore the only combinations of values for X, T, and E are the following:

X	T	E
1	1	0
0	0	0
1	0	1
0	1	-1

That in itself is not necessarily too hard to handle, but when it comes to the definition of E as "random", the definition of T as the average of parallel X's, and some of the assumptions of classical reliability theory such as X being normally distributed around T with constant standard error of measurement, things start to break down. Why? Think about it:

a. In order for E to be random (our first approach), the mean E must be equal to zero. Referring to the four possible combinations above, that is not a problem if all four combinations are equally represented in the data or if only the first two combinations are equally represented in the data and there are no 1,0,1 or 0,1,-1 combinations in the data at all (perfect reliability?), or if only the last two combinations are equally represented in the data and there are no 1,1,0 or 0,0,0 combinations in the data at all (perfect unreliability?). But many distributions of those combinations will yield a mean E other than zero.

b. In order for E to be random, the correlation between T and E must also be equal to zero. Again referring to the four combinations, if they are equally represented in the data the correlation between T and E must be negative, because there is no positive sum of cross-products of T and E to "balance out" the negative sum of cross-products produced by the fourth combination. (Do you follow that?)

c. What is a "parallel" item to "Who is the president of the United States?" or to the attitude-toward-abortion item?

d. Our Axiom #2 of classical reliability theory includes the assumption that for each person (I'll drop the "object" terminology since we're considering social measurement only) (her)his error scores are normally distributed around (her)his true score. You can't get a normal distribution of errors if they're all 1, 0, or -1.

<u>Some approaches to the estimation of the reliability of single items</u>

There are other problems also, but those are sufficient to suggest that we need a different kind of reliability theory to cope with single dichotomous items. One that has been invoked is the so-called "Platonic" theory where error scores are random and true scores are "deserved scores" but not "universe scores"--they are not averages across parallel forms. (For more on Platonic scores see Sutcliffe, 1965; Klein & Cleary, 1967, 1969; Bohrnstedt, 1983; and the summary by Traub, 1994 of the "sex of the chicken" example in Lord & Novick , 1968.)

The Knapp method (and comparison to the phi coefficient)

I adopted some of the principles of Platonic theory in an article I wrote several years ago about the reliability of a dichotomously-scored cognitive test item (Knapp, 1977b--an article in which the editors insisted that I use the non-sexist neologism "hir" instead of "her", "his", or "him"!). The principal findings of that article were as follows:

1. The "reliability for rights" is equal to $(1-d)^2$, where d is the proportion of "knowers" (people whose true score is equal to 1) who get distracted for whatever reason and give the wrong answer (and get an obtained score of 0). If none of them get distracted (unlikely, but not completely out of the ordinary), d =0 and the "reliability for rights" is equal to 1. If all of them get distracted (much more unlikely), the "reliability for rights" is equal to 0.

2. The "reliability for wrongs" is equal to $(1 - g/c)^2$, where g is the proportion of "non-knowers" (people whose true score is equal to 0) who guess randomly at the answer to the item and c is the number of choices provided in the item (c = 2 for a true/false item, for example), so that g/c of them give the right answer (and get an obtained score of 1). For an "open-ended" item (no choices provided) or for a multiple-choice item for which there is no guessing on the part of the "non-knowers", g/c = 0 and the "reliability for wrongs" is equal to 1. The "reliability for wrongs" can never be equal to 0.

The problem becomes one of determining the unknowns d and g (c will always be known). In my article I show how to estimate both d and g, using a test-retest approach (the only feasible empirical option), and consequently the "reliability for rights" and the "reliability for wrongs". (But if you think that the mathematics in this book is already too heavy for you, you may not want to try to follow the derivations!) I also go on to talk about weighting the "reliability for rights" and the "reliability for wrongs" to get an estimate of "over-all reliability" and to compare that value with the test-retest reliability calculated by applying the formula for the phi coefficient (the Pearson r for two dichotomies) to the following table of frequencies:

		Retest 1	Retest 0
Test	1	A	B
	0	C	D

$$\Phi = \frac{(AD - BC)}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

where A+B+C+D = 1

Unfortunately, the phi coefficient only provides an indication of the relative relationship between the test and the retest obtained scores, and if any of the sums in the denominator of the formula for Φ are equal to zero, Φ does not exist.

For the Knapp method, B and C should be approximately equal (if the test-retest occasions are “parallel” that should not be a problem) and A+D should be greater than B+C (if it’s not, you have a very unreliable item!)

It's time for an example. Consider one of the examples (a hypothetical four-choice multiple-choice item) I used in that 1977 article (I also included several real-data examples):

		Retest	
		1	0
Test	1	.40	.10
	0	.20	.30

For those data, ONE solution for d and g is d = .248 and g = .273. (Alas, that solution is not unique, but reasonable.) This suggests that about 25% of the "knowers" were distracted and gave the wrong answer and about 27% of the "non-knowers" guessed, with about one-fourth of the 27.3%, i.e., 6.8%, guessing correctly (since c = 4). Plugging the values for d, g, and c into the formulas for the "reliability for rights and the "reliability for wrongs" we get:

Rights: .566 (too many "knowers" were distracted)
Wrongs: .868 (better; there were not too many "lucky guessers")

There were estimated to be about 70% "knowers" (70.4% to one decimal place) and about 30% "non-knowers" (29.6% to one decimal place). If the "reliability for rights" is weighted by .704 and the "reliability for wrongs" is weighted by .296, one estimate for the "over-all reliability" of that item is .655.

The phi coefficient for the same table is .408, which suggests poorer reliability. But the .655 and the .408 are not directly comparable, since the Knapp reliability coefficient can only take on values between 0 and 1, whereas Φ can range between -1 and 1.

The Guttman method

My approach to the reliability of a single-item instrument and the phi-coefficient approach aren't the only possibilities. Several years earlier, Guttman (1946) provided a strategy for estimating lower bounds and upper bounds for the reliability of a single item that "works" for dichotomous cognitive items such as "Who is the president of the United States?" but is particularly appropriate for multi-categoried affective "Likert-type" items such as the attitude-toward-abortion

item. He showed (again I'll spare you the mathematical details) that if you only have empirical data for one administration of an item, the lower bound for the reliability of the item is given by (in notation different from his):

$(k/k-1)(f_{max} - 1/k)$, where k is the number of response categories and f_{max} is the largest relative frequency for any of the categories.

He gave as an example a three-categoried item (not unlike the abortion item) for which the relative frequencies are:

Yes: .60
Undecided: .15
No: .25

For those data, k = 3 and f_{max} = .60, so the lower bound is (3/2)(.60 - 1/3) or .40, which means that the reliability of the item is at least .40 (and, of course, at most 1). He went on to point out that had all of the relative frequencies been 1/3 each, the lower-bound would have been 0, which is also the case by the definition of a reliability coefficient, so a rectangular distribution of obtained scores for a single item is absolutely no help at all in estimating reliability by Guttman's method, if you have data for just one administration of the item.

Things are much better, as you might imagine, if you have test-retest data for two administrations of the item. He showed that the lower-bound and the upper-bound are functions of the largest relative "sub-frequency" for each row of the k-by-k contingency table ("cross-tab") for the two testing occasions, if the contingency table is approximately symmetric and the sum of the entries in the principal diagonal is greater than the sum of the off-diagonal entries (the same as the Knapp method assumptions) . The notation gets a bit cumbersome, so I'll use Guttman's test-retest example to explain the procedure. Here are the data:

		Retest		
		Yes	Undecided	No
Test	Yes	.10	.15	.05
	Undecided	.15	.25	.05
	No	.05	.05	.15

For those data, the largest relative frequencies in the rows are the .15 in the first row, the .25 in the second row, and the .15 in the third row. Calculate the sum of

those frequencies, .55 in this example, and call it ∑f. Then the lower-bound for the reliability of the item is (k/k-1)(∑f - 1/k), where k is the number of response categories, as before. (Note the similarity to the formula for the lower bound for one administration of the item.) For A = .55, this works out to be (3/2)(.55 - 1/3) or .33. Guttman (1946) gives two formulas for the upper-bound. The more precise formula is {k/(k-1)}{h - 1/k}, where h is equal to (1/k){1 + v[(k-1)(k∑q -1)]} and ∑q is the sum of the relative frequencies in the principal diagonal (upper left to lower right) of the contingency table. For the example, ∑q = .10 + .25 + .15 = .50, h = (1/3){1 +v[2((3)(.50) - 1)]} = 2/3, and the upper bound is therefore (3/2)(2/3 - 1/3) = 1/2 or .50. Since those lower and upper bounds are fairly "tight" (.33 and .50), we have a good fix on the reliability of the item even though we don't have a unique solution.

Percent agreement and Cohen's kappa

Another strategy that has been around for many years is the simple "proportion agreement" or "percentage agreement" method, which for the arithmetic item is .40 + .30 = .70, or 70%; and for the abortion item is .10 + .25 + .15 = .50, or 50%. (Those are the relative frequencies in the principal diagonal of the respective contingency tables.) See Wakefield (1980) for a comparison between percentage agreement and the phi coefficient.

And there is of course the popular variation on proportion agreement, Cohen's kappa (Cohen, 1960), which corrects the proportion agreement for any agreement that might be attributable to "chance". The formula for Cohen's kappa is:

$$\kappa = \frac{p - p_c}{1 - p_c}$$

where p is the actual proportion of agreement and p_c is the proportion of "chance" agreement. This statistic has been the subject of a huge and often controversial literature, and has been extended to more than two testings (usually ratings) and to polytomous scoring. (See, for example, Cohen, 1968; Fleiss, 1965, 1971, 1975, 1981; Brennan & Light, 1974; Hubert, 1977; Landis & Koch, 1977; Brennan & Prediger, 1981; Davies & Fleiss, 1982; Brook & Stirling, 1984; Lee & Suen, 1984; Darroch & McCloud, 1986; Topf, 1986; Maclure & Willett, 1987; Dunn, 1989; Stine, 1989; Feinstein & Cicchetti, 1990; Cicchetti & Feinstein, 1990; Brennan & Hays, 1992; Hutchinson, 1993; Byrt, Bishop, & Carlin, 1993; Knapp & Brown, 1995; Dunn, 2004). Some of the controversy revolves around the determination of p_c. For our hypothetical arithmetic item, p is .40 + .30 = .70, but what do we use for p_c ? The probability of chance success is 1/4 or .25 for either testing (random guessing without even reading the item), so the probability of chance success for both testings is (1/4)(1/4) = 1/16 or .0625 (assuming independence for the two eventualities). Substituting that for p_c we

would get κ = (.70 - .0625)/(1 - .0625) = .680. But that assumes that both "knowers" and "non-knowers" are equally prone to random guessing. I don't believe that; do you?
The situation is even worse for the abortion item. It's one thing to assume that there might be some random guessing for a cognitive item (by "knowers", "non-knowers", or both), and to want to correct for that. It's quite another thing to assume that when people are asked for their attitude about something they might guess at the various response options. And Cohen's kappa is actually more often applied to the estimation of inter-rater reliability where both raters use the same k-point scale, and p is again corrected for chance agreement. Chance agreement? Do competent raters ever rate randomly? If you think that they do, you can always demand the resolution of a higher value for p.

Spearman-Brown in reverse

Perhaps the most common, but the most dangerous (in my opinion) approach to the estimation of the reliability of a single item is to use the generalized Spearman-Brown formula (see following chapter) "in reverse". That formula was developed to provide a way of estimating what the reliability of a test k times as long as the one in hand would be, if similar items were added to the existing instrument. The same formula can be used to estimate the reliability of a test "1/k th" as long as the one in hand, so if the existing test has k items the formula would produce an estimate of the reliability of one (any one) of its items. Although there's nothing wrong with that mathematically, there are at least two problems substantively: (1) you already have the k-item test, so why do you even care about a test that is 1/k th as long?; and (2) the estimates that are yielded are often discouragingly small and (again in my opinion) are serious under-estimates of single-item reliability.

Visual analog(ue) scales

One of the most common single-item (but not dichotomous) measuring instruments is the visual analogue scale (VAS); the -ue ending on analogue is sometimes dropped. It is used primarily in self-reports of various perceptions such as pain, anxiety, and the like. The measuree is asked to indicate on a scale that is typically 10 centimeters in length the level of pain, anxiety, or whatever (s)he is experiencing at the present time. The scales have verbal descriptors at their opposite poles (e.g., “no pain” and “excruciating pain”); some also have additional descriptors throughout the scale (e.g., “minor annoyance”) and some have numbers associated with the descriptors. It can be presented either horizontally or vertically, and the measuree can use a pen or pencil to indicate (her) his perception or can call out the level to the measurer. The person’s score is the distance from the bottom of the scale to the indicated point (and is thus an attempt to “continuize” the typical Likert-type scale so that traditional descriptive and inferential statistics are more defensible).

An interesting recent example of a VAS is the Distress Thermometer (unfortunately abbreviated to DT) that is used to measure the level of distress that cancer patients are experiencing. (See Roth, Kornblith, et al.,1998; Jacobsen, Donovan, et al., 2005.) From the name of the instrument you can tell that it is presented vertically, with "no distress" (0) at the bottom of the scale, "extreme distress" (10) at the top of the scale, and with the intermediate numbers 1,2,3,4,5,6,7,8,9 spaced evenly along the left-hand side of the scale. The only scores that are reported are those integers from 0 to 10 (i.e., the scale is no further "continuized").

As far as the reliability of visual analogue scales is concerned, the choice of approach is usually limited to test-retest (measure-remeasure), but the time interval between test and retest is crucial, since the true perception can change dramatically over very short periods of time. Inter-rater and intra-rater approaches may be used if someone other than the person experiencing the pain, anxiety, etc. is indicating the level. In the first of two short pieces on the VAS that are available on the internet, Johnson (2005) even suggested the creation of two parallel forms by using polar descriptors such as "no pain" and "intense pain" for one form; and using "no pain at all" and "worst possible pain" for the other form. Whether those forms are truly parallel, and whether they would "work", remain to be seen.

For more on visual analogue scales, see Wewers and Lowe (1990) and Cline, Herman, et al. (1992).

Additional reading

Use of the generalized Spearman-Brown formula in reverse dates back at least as long ago as Holzinger's article (Holzinger, 1932) and is one of the reasons why single items have a bad reputation, in spite of the empirically-demonstrated high reliability (and validity!) for many of such items. For more on the reliability of single items, see Sprott and Vogel-Sprott (1987); Cunny and Perri, 1991; Youngblut and Casper (1993); and Renzo (2002, 2003). For a particularly thorough discussion of the reliability (AND validity) of methods for resolving discrepancies between raters for single-item instruments, see Johnson, et al. (2003).

CHAPTER 8: The internal consistency of multi-item tests

For many researchers in the social sciences, "internal consistency reliability" is the only kind there is. Cronbach's coefficient alpha, in particular--see below--is used to estimate the reliability of social science instruments more often than all of the other methods taken together. I would like to begin this chapter by exploring some of the historical and practical reasons why this is so.

A little history

Most of reliability theory started with Spearman (1904), who was concerned with the precision, accuracy, dependability (call it what you will) of psychological tests of various sorts. He was familiar with the notion of measuring and re-measuring with the same form or with comparable forms of an instrument, but he was also aware of some of the problems entailed with the taking of more than one measurement on the same persons, not the least of which is the usual assumption that the true score (he didn't use that exact term) must remain constant between the first and the second testings. He (Spearman, 1910), and at about the same time, Brown (1910), came up with the idea of administering one form of the test once, dividing the test in half (creating two pseudo-parallel half-forms a and b), scoring both halves, finding the correlation between the obtained scores on the two halves, and then "stepping up" that correlation in order to estimate what the correlation might have been between two parallel full-forms A and B. That estimate was

$$r_{AB} = \frac{2r_{ab}}{1 + r_{ab}}$$

For example, if the correlation between two half-forms of a test were .60, the estimated correlation between two full-forms would be 2(.60)/(1 + .60) = .75. That value of r_{AB} , .75, would then be taken to be an estimate of the reliability coefficient, r_{XX} , for the test.

The difference between r_{ab} and r_{AB} is sometimes surprisingly large and sometimes surprisingly small. Recently, Wainer and Thissen (1996), in their discussion of the reliability of "testlets" (sets of items all based upon a single stimulus), gave an example of the doubling of a test that increased its reliability from .85 to .92. Having to develop twice as many test items would appear to be a rather expensive price to pay for an increase in reliability of .07, unless the higher reliability was thought to be absolutely essential.

The Spearman-Brown formula has been generalized as follows:

$$r_{kk} = \frac{kr_{XX}}{1 + (k-1)r_{XX}}$$

where r_{XX} is the reliability of the test that is actually in hand and r_{kk} is an estimate of the reliability of the test if it were to be made k times as long (by adding similar items). Pursuing that same example, if a reliability coefficient of .75 were not acceptable you might consider making the test three times its new length, which would result in an estimated reliability of 3(.75)/ [(1 + 2(.75)] = 2.25/ 2.50 = .90.

In the previous chapter I bemoaned the fact that some people estimate the reliability of a single item by using the generalized Spearman-Brown formula "in reverse". That is, they know the reliability for their test of k items, and they'd like to know how reliable a test 1/kth as long would be (1/k times k is 1). Continuing with the present example, if the test with reliability of .75 had 20 items (10 odd-numbered and 10 even-numbered), substituting 1/k = 1/20 for k = 20, and .75 for r_{XX} , the single-item reliability is estimated to be (1/20)(.75)/[1 + (-19/20)(.75)] or approximately .13. That is of course a terribly low reliability--much too low to be believed, in my opinion. (For more on this, see Knapp & Brown, 1995.)

On the one hand, the "split-halves" idea was ingenious (saves a heck of a lot of work, for one thing!), but on the other hand it was also a bit weird (chopping in half and then re-creating the whole?). In any event, it was the method of choice for many years. Some people worried about how the test should be divided in half, but allocating the odd-numbered test items to one half-form and the even-numbered items to the other half-form soon became the universally accepted way to do so. (Tests with an odd number of items presented a minor problem, since the odd-numbered “half” would have one more item than the even-numbered “half”.)

Kuder and Richardson

Concern about the various ways of dividing a test in half (and the possibility of getting a different reliability estimate for each division) grew over the years. In 1937, Kuder and Richardson wrote a long article in which they derived several formulas for estimating the reliability of an instrument consisting of k dichotomously-scored items that could only be administered once. Starting with the notion of the correlation between a form that actually exists and a hypothetical parallel form that does not exist, and making successively more restrictive assumptions, they arrived at two formulas that became the most popular for estimating the reliability of a multi-item instrument. One of these, their Formula #20, which still goes by that name, is

$$r_{XX} = [k/(k-1)] [1 - \sum(p_i q_i) / S_X^2]$$

where p_i is the proportion of "measurees" who answer item i correctly (i = 1, 2, ..., k), $q_i = 1 - p_i$ is the proportion of "measurees" who answer the item incorrectly--so $p_i q_i$ is the obtained variance of item i, and S_X^2 is the obtained variance of the total-test scores.

The other formula, which came next in their derivations, Formula #21, which is also still called that, is

$$r_{XX} = [k/(k-1)] [1 - M_X (k - M_X)/ S_X^2]$$

where M_X is the mean of the obtained total-test scores (and is equal to $\sum p_i$).

The same Kuder-Richardson Formula #20 was later derived by Hoyt (1941), by Burt (1955), by Lord (1955), and by others, under slightly different sets of assumptions, using an analysis-of-variance approach. (Shoemaker, 1969 later showed that items answered correctly by all examinees or by no examinees can dramatically lower the value produced by Formula #20.) Formula #20 and Formula #21 both represented considerable improvements over Spearman and Brown's split-half technique, the former because it yielded a single estimate (although it involves more work) and the latter because it was so simple to use--all you need are the number of items, the total-test mean, and the total-test variance.

Cronbach

Two problems remained, however. The first was all of the assumptions that needed to be made, especially for Formula #21--that all of the test items are of approximately equal difficulty. This is almost never the case, and you can get very strange results if there is a considerable range in difficulty. In an appendix to my article on coefficient alpha (Knapp, 1991) I gave a hypothetical example of a set of four test items that constitutes a perfect Guttman scale when administered to a set of 16 people, and for which Kuder-Richardson Formula #21 is exactly equal to 0 (Kuder-Richardson Formula #20 equals .604), yet it is hard to imagine anything that is more reliable than a perfect Guttman scale. (For such scales if you know a person's total score you also know exactly which items (s)he answered correctly.) Therefore, WARNING: Don't use Formula #21 unless the items are very close in difficulty (for the example in Knapp, 1991 the item means ranged from .0625 to .9375).

The second problem was that the formulas only "worked" for dichotomously-scored items. In his now-classic (and actually very readable) article, Cronbach (1951) addressed both of those problems (and then some) and derived the following formula:

$$r_{XX} = [k/(k-1)] [1 - \sum(S_i^2)/ S_X^2]$$

where S_i^2 is the obtained variance for item i, whether it is dichotomously-scored or not. This is the generalization of Formula #20, and Cronbach gave it the symbol □ (the Greek alpha), so it is commonly referred to as "Cronbach's alpha" or "Coefficient alpha", or--by McDonald (1999)--"Guttman-Cronbach alpha" (because Guttman had previously derived it in his 1945 article as one of several lower bounds to an instrument's reliability--see Callender & Osburn, 1979 for a comparison of those lower bounds). The name alpha and the symbol □ are actually very unfortunate choices, because the same name and symbol are used for two other quantities in statistics: (1) the probability of making a Type I error (the "level of significance"); and (2) the Y-intercept for a population regression line, plane, or hyperplane.) There is also a "Formula #21 version" of Cronbach's alpha (but see Kuder, 1991, regarding what formulas should be given what names).

Various approximations to Cronbach's alpha have been derived. For example, if all of the item variances can be assumed to be approximately equal, the formula can be written as

$$r_{XX} = \frac{k\, r_{avg}}{1 + (k-1)\, r_{avg}}$$

where r_{avg} is the average (mean) of the correlations between score on item i and score on item j (I,,j = 1, 2, ...,k; i<j), i.e., the entries in the upper (or lower) triangle of the k-by-k matrix of inter-item correlations. (Edgerton & Toops, 1928, developed a handy method for calculating r_{avg} without actually calculating any of the r_{ij} .) That is the formula for what is called "standardized alpha" in the measurement literature and in computer packages such as SPSS and SAS.

Consider the following set of hypothetical data for a four-item test administered to five subjects. The data were originally provided by Kerlinger (1976) and were repeated by me (Knapp, 1991).

	Item 1	Item 2	Item 3	Item 4
Subject 1	6	4	5	1
Subject 2	4	1	5	4
Subject 3	4	6	4	2
Subject 4	3	6	4	3
Subject 5	1	2	1	2

For those data, alpha is .449 (try it) and "standardized alpha" is .404.

Cronbach proved that his alpha is approximately equal to the average of the "stepped-up" Spearman-Brown reliabilities for all possible ways of dividing a k-item test into two equal halves. (There are $k! / 2[(k/2)!]^2$ such ways, if k is even.) If you don't believe Cronbach's claim, try it out on the above data. Put items 1&2 in one half and items 3&4 in the other half, find the correlation r_{ab} between those two half-tests, "step it up" by calculating $r_{AB} = 2r_{ab} /(1 + r_{ab})$, record that value of r_{AB} ; then repeat the process two more times (1&3 vs. 2&4; 1&4 vs. 2&3) and find the average of those three values of r_{AB} . (The number of possible ways for dividing a four-item test into two equal halves is three.) Alpha will be exactly equal to the average of all the split-half reliabilities whenever all of the half-tests have the same variances. That is very important. As Novick and Lewis (1967) pointed out, Cronbach's alpha is identical to the average of the split-half reliabilities determined by using Rulon's (1939) formula, not by using the traditional formula $2r_{ab}/(1+r_{ab})$.

Although Cronbach's coefficient alpha enjoys great popularity and its literature is vast, it has some drawbacks. For one thing, as I and others have shown (see, for example, Knapp, 1991; Krus & Helmstadter, 1993), it can actually take on any value between "minus infinity" and +1 (watch what happens if you switch the data for Item 3 from 5, 5, 4, 4, and 1 to 1, 4, 4, 5, and 5), even though a reliability coefficient is defined in such a way as to restrict its theoretical range from 0 to 1. For another thing, you can get unusually high alphas by having lots of items that correlate only slightly with one another. If, for example, r_{avg} is equal to .10 in the formula for standardized alpha, and if there are 100 items on the test, i.e., k = 100, alpha turns out to be a surprisingly large .92 (a large k has "swamped" the small r_{avg}). Those 100 items aren't really very internally consistent with one another, are they? (You can also get a .92 for 10 items that have an average inter-item correlation of .53. That sounds like the better instrument, all other things being equal, despite the considerably smaller number of items.)

How many items?

This leads nicely into a discussion of test length. How many items should a test have? If you want to have a high Cronbach's alpha, the answer is "lots" (but see above). Several authors have also studied the relationship between the number of items on a test and the standard error of measurement for the total score on the test. The general consensus (see Lord, 1957a, 1959a; Swineford, 1959; Gardner, 1970) is that tests of the same length have nearly identical standard errors of measurement. If the items are dichotomously scored, those scores are added together to get a total-test score, and the Kuder-Richardson Formula #20 is used to estimate the test's reliability, the standard error of measurement for the total-test scores can be closely approximated by $.43\sqrt{k}$, where k is the number of items on the test. (For the example just cited, with k = 100, the standard error of measurement would be estimated to be 4.3.)

Factor analysis and internal consistency reliability

This in turn leads nicely into a discussion of the connection between the reliability of a multi-item test and the (exploratory) factor analysis of the data obtained by administering such a test to a large group of subjects. (See Nunnally & Bernstein, 1994, or any other measurement theory textbook, for a discussion of the basic principles of factor analysis.) Since we have been talking all along in this chapter about the internal consistency of a set of test items, it would seem that if Cronbach's alpha is high (say .90 or above) a factor analysis should yield one big factor, suggesting that the items all "hang together" to measure "the same thing" (whatever the "thing" is--that's a validity problem). Well, yes and no. If all of the items inter-correlate highly with one another, say .7 and above, that will indeed be the case; i.e., you'll get one big factor and a bunch of little ones. As we have just seen, however, if the correlations between pairs of items are rather small, but there are lots of them, that will not be the case. The situation is actually very complicated. For the big k, small r_{avg} case just cited, if all of the inter-item correlations are approximately equal to one another, we'll get just one "eigenvalue" that is greater than one, but we'll also get 99 eigenvalues that are just a little bit less than one. (Eigenvalues are very important in factor analysis, especially those that are greater than one. See the article by Joe & Mendoza, 1989 on “The internal correlation”--and the comments regarding that article in the same issue--for further discussion of the connections between the eigenvalues of a correlation matrix and internal consistency reliability.) If the inter-item r's are not all equal but "average out" to .10 (with some medium-size correlations of, say, .50 and some very small and/or negative correlations to "balance" them), and with k still large, we'll get a multi-factor solution. (See the Appendix to Carmines & Zeller's 1979 monograph on reliability and validity for a good summary of the "tie-in" between factor analysis and internal consistency reliability.)

Kaiser (1960) claimed that any factor for which the associated eigenvalue is less than one would have negative reliability (there we are with negative reliability again!) and would therefore be essentially useless as a subscale variable. (See also LaForge, 1965). Cliff (1988) disputed Kaiser's claim, arguing that factors with eigenvalues less than one could have positive reliability. Kaiser (1991) later justified his original claim, and Cliff and Caruso (1998) reiterated Cliff's contention. That little-known controversy has yet to be completely resolved (Li & Wainer, 1997 support Kaiser) and it would appear to be very difficult to do so, because it involves a number of assumptions regarding how the reliabilities are to be estimated, what method of factor analysis is employed, etc. In case you didn't already know, factor analysis is a very tricky business!

For more on the relationship between factor analysis and the internal consistency reliability of a measuring instrument, see Wherry and Gaylord (1943), McDonald (1999), and the discussion of that relationship on the statsdirect.com website.

Inter-item correlations and Item-to-total correlations

Standardized alpha provides a direct indication of the extent to which the items inter-correlate with one another. A similar indicator is provided by the correlations between each of the test items and the total-test score (the higher such correlations are, the greater the internal consistency). The formula for Cronbach's alpha can be written in a way that involves both (see Gulliksen, 1945 and Ebel, 1967 for other formulas):

$$r_{XX} = [\, k/(k-1)] \, [\, 1 - \sum r_{Xi} S_X S_I - \sum r_{ij} S_I S_j) / (\sum r_{Xi} S_X S_I) \,] \,,$$

where r_{Xi} is the correlation between obtained total-test score and score on item i (i = 1, 2, …, k), r_{ij} is the correlation between score on item i and score on item j (i, j = 1, 2, …, k), and the other symbols are as previously defined. There is a problem regarding whether to include or to exclude the score for the item itself as part of the total-test score when calculating the correlation between item and total (it is included in the given formula), but that is a relatively minor consideration when there is a large number of items.

Other approaches to internal consistency reliability

There are other measures of the internal consistency of multi-item tests, and Scott (1960) summarized several of them in his article entitled "Measures of test homogeneity". ("Homogeneity" is sometimes used as a synonym for "internal consistency", as far as reliability is concerned.) In addition to those that Scott discusses, there is Webster's (1960) generalization of Kuder-Richardson Formula #21; Heise and Bohrnstedt's (1970) omega (McDonald, 1999 has a different omega); Armor's (1974) theta; Raju's (1977) beta; and many more--see Greene and Carmines (1980) for a comparison of some of them, and see Osburn (2000) for several others. As you can see, measurement experts like to use Greek letters to designate their reliability coefficients. I guess it's all Greek to them!

Inter-rater and intra-rater reliability

In my opinion, all of the preceding “internal consistency” approaches to assessing the reliability of the total score on a test can be applied to assessing inter-rater and intra-rater reliability. Just make the examinees “ratees” and the items “raters” and you can determine the reliability of the sum (or average, i.e., arithmetic mean) of the ratings given to the people being rated, and it won’t matter whether the raters are different judges rating each person once or it’s the same judge giving multiple ratings to each person. Substantively it’s a different problem but methodologically it’s the same thing.

Or is it? There was an interesting controversy aired in the pages of Personnel Psychology a few years ago, between Frank Schmidt and his colleagues on one side and Kevin Murphy and his co-author on the other side. Murphy and DeShon

(2000) took strong exception to the claim made by Schmidt and Hunter (1996), and by Viswesvaran, Ones, and Schmidt (2000), that ratings and items are methodologically interchangeable, arguing that correlations between ratings did not provide appropriate evidence for either inter-rater reliability or intra-rater reliability. In their counter-argument, Schmidt, Viswesvaran, and Ones (2000) reiterated their position even more strongly. The controversy actually spilled over into matters involving the correction for attenuation (see Chapter 4, above) and generalizability theory (see Chapter 12, below)! [Did you think that research methodologists all agreed with one another regarding data analysis? If so, think again. Frank Schmidt, for example, seemed to make a habit of getting himself involved in such controversies, the most well-known being his views regarding the usefulness (or, in his opinion, the non-usefulness) of significance testing in psychological research. See Schmidt (1996).]

Additional reading

There have been a number of other contributions, theoretical and empirical, to the literature on internal consistency reliability. To name most of them, there are the articles by Brownell (1933), Read (1939), Himes (1989), Charter (2001), and Feldt and Charter (2003) regarding split-halves; Dressel's (1940), Tucker's (1949), Zimmerman's (1972), Cudeck's (1980), and Cliff's (1984) work on Kuder-Richardson Formula #20; Saupe's (1966) suggestions for selecting items that will increase the reliability of simple change scores; Maberly's (1967) investigation of internal consistency within particular ranges of obtained scores; Krus and Helmstadter's (1987) reformulation of the generalized Spearman-Brown formula; Cronbach, Schonemann, and McKie's (1965) article on coefficient alpha for stratified parallel tests; Cronbach's (1988) general discussion of the internal consistency of tests and his (and Shavelson's, 2004) later reflections on alpha (published after Cronbach's death); Green, Lissitz, and Mulaik (1977), Cortina (1993), Schmitt (1996), Rogers, Schmitt, and Mullins (2002), and Green (2004) on alpha; Ferketich's (1990) comparison of alpha, omega, and Armor's theta, using a real-data example from nursing research; Berk's (2000) humorous article concerning questions and answers about K-R 20; Feldt & Charter's (2006) article on averaging internal consistency reliability coefficients; and Rae's (2006) article on correcting alpha when measurement errors are correlated.

For everything you want to know about inter-rater reliability, see LeBreton, et al. (2003). Also see Stemler (2004) for an interesting discussion of three kinds of inter-rater reliability.

CHAPTER 9: Intraclass correlations

When anyone mentions a correlation coefficient, what immediately comes to mind is a Pearson product-moment correlation coefficient, the traditional indicator of the direction and the magnitude of the linear relationship between two variables. In this book I have had a great deal to say about Pearson r's as estimators of reliability coefficients, as affected by attenuation, etc. It might surprise you to know that the product-moment correlation coefficient has a relatively brief history (about a hundred years), having been developed by Karl Pearson at the beginning of the 20th century (Pearson, 1904) at the urging of his friend and colleague Francis Galton (yes, that Galton), who was Charles Darwin's cousin (yes, that Darwin!). And there is another correlation coefficient, the intraclass correlation coefficient (ICC) that has a similar history. It was also developed by Pearson (some people claim it was Harris, 1913 and not Pearson) and refined by his nemesis, R.A. Fisher (yes, that Fisher!!), who called Pearson's "other" correlation (the now- traditional one) the interclass correlation coefficient .

There are a number of technical similarities and differences between the two, but for our purposes here the principal ones are: (1) most applications of intraclass correlations are to within-variable problems and most applications of interclass correlations are to between-variable problems; and (2) mathematically (but not necessarily substantively) the two are identical if the variances of the measurements being correlated are identical. (See Robinson, 1957 for more on this.)

There is good news and bad news regarding intraclass correlations. The bad news is that there are ten different kinds (see McGraw & Wong, 1996). The good news is that we'll only consider two of them in this chapter.

The most useful one

Let's start with an example. Suppose you were interested in the test-retest reliability of a measuring instrument and you have measured n people k times each, but you have no interest in, and no way to sort out, which measurement was the first one, which measurement was the second one, etc. You display the data in a table where the rows of the table are the individual people and the columns are the measurement occasions, but it doesn't matter for a given individual which measurement goes in which column (the columns are not distinguishable). All that matters is the variability within person and the variability between persons. The formula is (see David Howell's discussion of intraclass correlation for unordered pairs at the following website: uvm.edu/~dhowell/StatPages/More_Stuff/icc/icc.html)

$$ICC = \frac{(MS_{br} - MS_{wr})}{MS_{br} + (k-1)MS_{wr}}$$

where MS_{br} and MS_{wr} are the mean square between rows and the mean square within rows, respectively; and k is the number of columns (here k = 2).
The mean squares are calculated as follows:

$MS_{br} = \{k[\Sigma (M_i - M_g)^2)]\} / (n-1)$, where M_i is the mean for row i (i = 1,2,…, n) and M_g is the grand mean

$MS_{wr} = \{\Sigma\Sigma (X_{ij} - M_j)^2\} / n(k-1)$, where X_{ij} is the observation in row i and column j (i = 1,2,…, n; j = 1,2,…, k)

n-1 and n(k-1) are the numbers of degrees of freedom for between rows and within rows, respectively.

If you're comfortable with the analysis of variance, consider our hypothetical example in Chapter 3, keeping the row designations as individual persons and the “First X” and “Second X” columns as the two height measurements. Calculate the ICC for the data as given. You should get .556. Then shuffle two or three of the within-row heights from one column to the other and calculate the ICC for that layout. You should get .556 again. And no matter how many within-row interchanges you make you’ll always get .556. (The Pearson interclass correlation will keep changing.)

In their article on intraclass correlations where the rows are “ratees” and the columns are “raters”, Shrout and Fleiss (1979) call this particular intraclass correlation ICC (1,1). It is the coefficient of choice for the situation in which the objects (usually people) being measured have been randomly drawn from a population (or can be “regarded” as such) and the k measurements taken on each of the objects do not have a distinguishable order, so that they also can be “regarded” as random.

The one that’s equal to Cronbach’s alpha

But what if the measurements are distinguishable and “fixed” (rather than random), and you are interested not in the reliability of an individual observation but in the reliability of the sum of the k observations for an individual (such as the total score on a test of k items). In that case there is a different intraclass correlation that is appropriate, which Shrout and Fleiss called ICC (3,k). Its formula is different from the formula for ICC (1,1), and so is its interpretation. Here’s the formula:

$$ICC(3,k) = \frac{MS_{br} - MS_e)}{MS_{br}}$$

where MS_{br} is as before and MS_e is the "error" or "residual" mean square.

The latter mean square is calculated as follows:

$$MS_e = \Sigma (X_{ij} - M_i - M_j + M_g)^2$$

where M_i is the mean of column i (i = 1,2,…k) and the other symbols are as before.

Consider the data that were used to illustrate Cronbach's alpha in the previous chapter. For that layout, n = 5, k = 4, and ICC (3,k) = .447. We got .449 for Cronbach's alpha for those data, using a different formula (and rounding). Coincidence? No; it can be shown (I won't do it, but Bravo & Potvin, 1991 did) that ICC (3,k) and alpha always produce identical results. As a matter of fact, Hoyt (1941) used the formula for ICC (3,k) [it wasn't called that at the time] when he re-derived the Kuder-Richardson Formula #20.

ICC (3,k) is therefore the intraclass correlation of choice whenever the columns are fixed, the rows are random, and the focus is on the reliability of a sum. That particular intraclass correlation is quite commonly employed in inter-rater reliability investigations (see Laschinger,1992 for an interesting example in nursing research) where the rows are "ratees" and the columns are "raters". But for such investigations the researcher is often interested in both the reliability of a "typical" rater and one for the reliability of the rater consensus (summed or averaged across raters). The former is called ICC (3,1) [see McGraw & Wong, 1996 for its formula--it's one of their ten and one of Shrout & Fleiss's six]; the latter is the ICC (3,k) with which we have just been concerned.

If the ratings are 1-to-n rankings of n objects by k raters, Spearman's rank correlation coefficient or Kendall's tau are to be preferred to intraclass correlations for two raters, with Kendall's coefficient of concordance being appropriate for more than two raters.

Intraclass correlations also play a dominant role in the determination of the generalizability coefficient analogues of classical reliability coefficients and in the resolution of various unit-of-analysis problems (see Chapter 12).

If you'd like to have an intraclass correlation calculated, there is a marvelous website, sip.medizin.uni-ulm.de/informatik/projekte/Odds/icc.html, that will calculate it (actually six of them) for you. All you need to do is enter the raw data by rows (or import them from a data file) and click the "calculate" button. Richard

Lowry's website, faculty.vassar.edu/~lowry/VassarStats.html, also will calculate ICC (1,1) for you and includes an excellent discussion of that concept.

Additional reading

For more on intraclass correlations and reliability I recommend Chapter 7 in Fisher (1925--it's a classic and is also very readable); three articles by Bartko and his colleagues (Bartko, 1966; Bartko, 1976; Bartko & Carpenter, 1976); the articles by Fleiss and Shrout (1978), by Armstrong (1981), by Johnson and Mott (2001), by Yen and Lo (2002); and the discussion in Dunn (2004).

CHAPTER 10: Two vexing problems

John reads four essays written by students A, B, C, and D, and assigns ratings of 1, 2, 4, and 5, respectively, to those essays. Mary reads those same essays and assigns ratings of 1, 3, 7, and 9, again respectively, to those essays. Is that evidence reflective of good or bad reliability? In what follows, that question will be answered by considering two different aspects of scientific measurement: (1) the matter of relative vs. absolute agreement; and (2) the matter of ordinal vs. interval measurement.

Relative vs. absolute agreement

In most of the previous chapters (Chapter 7 was the exception) it was taken for granted that obtained scores (as well as the corresponding true scores) were continuous and the principal concern was the relationship between two operationalizations of the same construct (parallel forms or measurement and re-measurement with the same form). Relationships in the form of correlations between two variables, especially Pearson product-moment correlation coefficients, tell you something about relative agreement but nothing about absolute agreement. When should you care about relative agreement and when should you care about absolute agreement? Consider as an example the data given above. There is perfect relative agreement between John and Mary. If we call John's ratings X and Mary's ratings Y, the equation Y = 2X - 1 is satisfied for every student A, B, C, and D (do the math). But the absolute agreement is quite bad. The only student for whom their ratings are the same is Student A; and their ratings for Student D differ by four points. (Mary's ratings are higher for every student other than Student A.) Hmmm.

The matter of absolute consistency vs. relative consistency has been a controversy of long standing. It was the basis for the disagreement between Lincoln (1932; 1933) and his critics (Franzen & Derryberry, 1932a; Ackerson, 1933) and "the jury is still out" regarding whether the emphasis should be placed on one, on the other, or on both. Engstrom (1988) provided some particularly good examples of various combinations of high and low absolute and relative agreement.

Mean and median absolute differences

Two simple indicators of the "typical" absolute measurement error are the mean and the median of the absolute differences between paired measurements, with the mean to be preferred if the magnitudes of all of the discrepancies are to be taken into equal account, but with the median to be preferred if you want to minimize the effect of any "outliers". For our hypothetical height data the mean absolute difference is 4.57 and the median absolute difference is 4. For the real flow data the mean absolute differences are 14.71 for the Wright meter and

19.35 for the Mini Wright meter, and the median absolute differences are 8 and 13, respectively.

Although there is a direct algebraic connection between the reliability coefficient and the standard error of measurement, namely $S_E = S_X \square (1 - r_{XX})$, there is no such connection between the reliability coefficient and either the mean absolute difference or the median absolute difference.

Ordinal vs. interval measurement

There is a type of variable called an “ordinal scale”, regarding which there has been even more controversy than for difference scores or Cohen's kappa! "Liberal" researchers treat such scales as though they are just like continuous or near-continuous "interval scales", whereas "conservative" researchers complain loudly "You can't do that!" This controversy began with the publication of S.S. Stevens' article that provided the nominal, ordinal, interval, and ratio taxonomy (Stevens, 1946) and continues to the present day. [I made two attempts at trying to resolve the controversy--Knapp, 1990 and 1993--to no avail.] If you are willing to treat those k-point (where k = 3 or more) "Likert-type" scales as interval scales, then obtained scores, true scores, and error scores can be added and subtracted with reckless abandon. If you are not (and I am not) you have two choices: "deflate" them to nominal status and use various extensions of Cohen's kappa (for example), or use statistical methods that have been "tailor-made" for ordinal scales (see, for example, Agresti, 1984). The latter choice is by far the more defensible one, but the price you must pay is learning a new set of formulas and procedures.

In Chapter 7 I described the approach taken by Guttman (1946) to estimating the reliability of an ordinal-level variable. The example chosen to illustrate that approach was a three-categoried variable (yes, undecided, no) with the relevant data arrayed in a 3x3 contingency table. Guttman’s formulas provided only a lower bound and an upper bound to the reliability for ordinal scales. There have been several other attempts to treat the reliability of ordinal-level variables as special cases of the general relationship between any two such scales. A brief discussion of three of them now follows.

Kendall’s tau-b

In their classic text on rank correlation methods, Kendall and Gibbons (1990) [the fifth edition of Maurice Kendall’s book; he died in 1983] described a procedure for handling the relationship between two ordinal variables that need not have the same number of scale points. They suggest arraying the cross-tabulated frequency data in an rxc contingency table, where r is the number of scale points for one of the variables and c is the number of scale points for the other variable. Kendall’s tau-b, a function of the extent to which pairs of observations are in the same or different orders for the two variables, provides an indicator of the relative

agreement between the two variables. As far as reliability is concerned, that same tau-b can be used as an indicator of the reliability coefficient for an ordinal variable (for a test/re-test situation, for example), where the number of rows and the number of columns of the associated contingency table is the same.

Goodman & Kruskal's gamma

An approach similar to that of Kendall was taken by Goodman and Kruskal (1984), resulting in a slightly different indicator of the relationship between two ordinal variables, but based upon the same concept of concordant or discordant pairs of observations. (See their book or Agresti's 1984 book for details.)

Williams' method

My favorite method for estimating the reliability of an ordinal variable is an application of canonical correlation analysis to a kxk contingency table of frequencies. In my 1993 article "Treating ordinal scales as ordinal scales" and in my article on contingency tables (Knapp, 1999) I credited Williams (1952) with the origination of this method, but there is an alternative history involving Maxwell (1961) [see Chapter IV, especially pages 69-72, of that book] , Marascuilo and McSweeney (1977), and Marascuilo and Levin (1983). I would now like to apply Williams' method to the data in the 3x3 table used to illustrate Guttman's (1946) method.

Here are the data (repeated from Chapter 7):

		Retest		
		Yes	Undecided	No
	Yes	.10	.15	.05
Test	Undecided	.15	.25	.05
	No	.05	.05	.15

You need actual frequencies (rather than proportions of the total group size) in order to use Williams' method. Guttman didn't provide the frequencies, so for convenience let's use a total group size of 200 (the actual number won't matter), so that the revised table is as follows:

		Retest			
		Yes	Undecided	No	∑
Test	Yes	20	30	10	60
	Undecided	30	50	10	90
	No	10	10	30	50
	∑	60	90	50	200

The method involves the creation of certain matrices and vectors, the calculation of the eigenvalues and the eigenvectors of one of those matrices (see Knapp, 1993, Appendix B, for the details), and the derivation of "scores" for the row and column designations. The indicator of reliability is the square root of the second-largest eigenvalue, which for these data is .4736 (not great; there are lots of off-diagonal frequencies). That number is between the lower bound of .33 and the upper bound of .50 determined by Guttman's method in Chapter 7, so all is well.

Back to John and Mary

If we had more essay-rating data for graders John and Mary, and if we could defensively treat the rating scale as an interval scale, a Pearson product-moment correlation coefficient for the relationship between John's ratings (X) and Mary's ratings (Y) would be perfectly fine for providing information regarding the (relative agreement) inter-rater reliability of that scale. If absolute agreement were of concern, either the mean or the median of the absolute differences of corresponding X and Y values should do the trick.

If, on the other hand, the rating scale is to be treated as an ordinal scale, then Kendall's tau-b, Goodman & Kruskal's gamma, or Williams' canonical correlation coefficient is to be preferred. My personal vote would go to Williams.

Additional reading

For more on absolute vs. relative agreement I recommend the general discussions by Bruton, Conway, and Holgate (2000), by Rogosa (2002), and by Baker and Kramer (2003); and the example given by Labouvie, Bates, and Pandina (1997) regarding the test/retest reliability of an instrument devised to measure retrospective perception of first use of alcohol and drugs. For more on ordinal vs. interval measurement I recommend Cliff (1979), Marcus-Roberts and

Roberts (1987), Cliff and Keats (2003), and Biswas (2006).

CHAPTER 11: Statistical inferences regarding reliability

In the previous chapters we have assumed, implicitly or explicitly, that we had data for entire populations. There were several references to "a very large number of objects", although most of the time, for illustrative purposes and to keep things simple, the number of objects was actually not very large. And in Chapter 5 we did talk a little about inferences for individual true scores, but the matter of sampling and sampling error were generally ignored.

In this chapter we are going to "bite the bullet" and face up to the fact that in most scientific research regarding reliability coefficients, standard errors of measurement, and the like, we have a sample of objects that are measured once or twice or, if we're fortunate, several times (see, for example, the matter of growth curves alluded to in Chapter 6), and we want to estimate, or test hypotheses about, the corresponding values in the larger population of objects from which that sample has been drawn. (In formal parlance, we have statistics and we want to say something about parameters.) How we go about doing it will turn out not to be very easy but nevertheless tractable.

Parallel forms reliability coefficients

Let's take the easiest cases first. Suppose we have a sample reliability coefficient determined by correlating obtained scores on two parallel forms, e.g., the r_{AB} of .50 for the hypothetical height data introduced in Chapter 3. (In that chapter the seven people were treated as a population of basketball players; now they're assumed to be a random sample from a large population of basketball players.) r_{AB} is a special case of a Pearson product-moment correlation coefficient, and the sampling theory for Pearson r's is well-known (but a bit tricky).

Consider first the matter of point estimation, i.e., the determination of "the best" single estimate of the population correlation coefficient ρ_{AB}. Because Pearson r's are "boxed in" between -1 and +1, their sampling distribution is not symmetric (much less normal) unless the population correlation is equal to zero. Furthermore, unlike a variance but like a standard deviation, there is no simple "unbiased" estimate for a population correlation. About "the best" we can do is use the sample correlation itself (in our example, .50) as an estimate of the population correlation.

Interval estimation is more promising and also generally more informative (see Fan & Thompson, 2001), but here we must transform the sample r into something called "Fisher's z" (not to be confused with standardized variable z), construct a confidence interval around it, and then transform the endpoints of that interval back into r's. For our example, the Fisher's z that corresponds to an r of .50 is .55 (trust me); the standard error (sampling error, not measurement error--

that's important!) of a sample Fisher's z is approximately equal to 1/□(n -3), which is .50 for our example; and the 95% confidence interval for the population Fisher's z therefore extends from .55 - 2(.50) to .55 + 2(.50), i.e., from -.45 to 1.55. Transforming back (from Fisher's z to r), the 95% confidence interval for □ extends from -.42 to .92. That's a very wide confidence interval and doesn't provide a very precise estimate of ρ_{AB}, but the sample size is only 7, so you get what you pay for! (If you're not familiar with Fisher's z, see any good introductory statistics text.)

[Note in the preceding paragraph the switch in notation from N to n. The former is preferred for the number of observations in a population; the latter is preferred for the number of observations in a sample.]

When it comes to hypothesis testing there is another wrinkle. For most applications of hypothesis testing we are interested in determining whether or not a particular sample statistic is "significantly different from zero", so we test the null hypothesis that the corresponding population parameter is equal to zero and we hope (usually) that we are able to reject that hypothesis. Not so for reliability coefficients. A sample reliability coefficient could be statistically significantly different from zero but the sample size could be so large that the statistic could reflect a very unreliable instrument. A sample r_{AB} of .10 based on a sample of size 900, for example, is statistically significant at the .01 level, but is a very weak reliability coefficient by anyone's standards. Testing the null hypothesis that the population reliability coefficient is equal to zero is an example of what Abelson (1995, 1997) has variously called a “silly” or a “gratuitous” significance test.

But all is not lost. Fortunately, Fisher's z comes to the rescue again. Instead of testing the null hypothesis that the population parallel-forms reliability coefficient is equal to zero, we can test the "null" hypothesis that the population reliability coefficient is some other number, perhaps a reliability of .80. (Null hypotheses are testable hypotheses; they don't always have to have zero in them.) It could be that an established instrument is known to have .80 reliability and you would like to determine if your instrument is competitive. For our r_{AB} of .50, if we want to test the hypothesis that our sample of seven people could have come from a population in which the reliability coefficient is .80, we proceed as follows.

Null hypothesis: $\rho = .80$ (a Fisher's z of .1.10)
Alternative hypothesis: $\rho < .80$ (a one-tailed test this time, since what is at stake is whether or not our test is up to standard)

Significance level: .025 (to be conventional for a one-tailed test that corresponds to a two-sided 95% confidence interval--if you follow that!)

Test statistic: $\dfrac{.55 - 1.10}{.50}$ [the standard error is the same]

= -.55/.50 = -1.10

That test statistic is distributed approximately as a standardized normal variate, and we accordingly cannot reject the null hypothesis (the absolute value of -1.10 is less than the "critical value" of 1.65), i.e., our instrument could be competitive (and it might also not be!).

The interval estimation approach actually subsumes the hypothesis testing approach here (just as it did in Chapter 5). Since .80 is within the 95% confidence interval, it cannot be rejected as a hypothesized value for □.

Test-retest reliability coefficients

Statistical inferences concerning same-form test-retest correlations proceed exactly in the same way as those for parallel forms, since they are also special cases of Pearson r's, but it is important to keep in mind what has been said before, by Kelley (1923) and others, that same-form reliability coefficients may violate one or more of the basic tenets of classical reliability theory. (A special case of a correlation coefficient is one that has been corrected for attenuation--see Chapter 4. Jackson, 1942 and Hakstian, Schroeder, & Rogers, 1988 have provided procedures for making inferences regarding such correlations.)

If you're interested not merely in an inference from a sample parallel-form or test-retest correlation to a population parallel-form or test-retest correlation, but would like to estimate or test a hypothesis about the difference between two parallel-form or test-retest correlations (for either independent samples or dependent samples), formulas derived from extensions of Fisher's z sampling theory are available (again see any good statistics text that discusses such matters).

Intraclass correlations

Now for the tough stuff. Since all intraclass correlations are functions of mean squares, the analysis of variance involving those mean squares can be used to test the statistical significance of any particular intraclass correlation coefficient (if you use the right mean squares!). But as indicated above, such a test is usually silly. It is the estimation of a confidence interval for the population intraclass correlation that should be of principal interest. And in this section we shall concentrate on interval estimation for ICC (1,1). For a comprehensive discussion of the appropriate formulas for constructing confidence intervals for intraclass correlations of various kinds, see McGraw and Wong (1996).

A confidence interval for ICC (1,1) can be determined as follows:

Lower limit: $(F_L - 1)/[F_L + (k-1)]$, where $F_L = (MS_{br} / MS_{wr}) / F_{tabled}$
and where F_{tabled} is the value in the F table for the desired degree of confidence for the appropriate number of degrees

of freedom [which for ICC(1,1) are (n-1) and n(k-1)]

Upper limit: $(F_U - 1)/[F_U + (k-1)]$, where $F_U = (MS_{br}/MS_{wr})(F_{tabled})$

For the height example, n = 7, k = 2, $MS_{br} = 56$, and $MS_{wr} = 16$. For 6 degrees of freedom "across the top" and 7 degrees of freedom "down the side", the F_{tabled} value for 95% confidence is 3.87. Working out the arithmetic, the 95% confidence interval for the population ICC (1,1) would be from -.106 to .926. That's a terribly wide confidence interval, and the lower limit is negative (that can happen--see Chapter 8), but that's because the number of rows (people) is so small.

A confidence interval for ICC (3,k) can be determined in a similar fashion. In the next section we will do so, as Cronbach's alpha. [Recall that ICC (3,k) is identical to Cronbach's alpha.]

Cronbach's alpha

As I pointed out in Chapter 8, most social science researchers determine the reliability of their instruments by using the formula for Cronbach's coefficient alpha (or Kuder-Richardson Formula #20, which is its special case for dichotomous items). Feldt (1965, 1969, 1980) derived approximate sampling distributions for alpha itself, the difference between two alphas for independent samples, and the difference between the alpha for one instrument and the alpha for another instrument administered to the same subjects. Only the first of these will be treated here. (See also Payne & Anderson, 1968; Cleary & Linn, 1969a; Hakstian & Whalen, 1976; Kraemer, 1981; Woodruff & Feldt, 1986; Feldt , Woodruff, & Salih, 1987; Mendoza, Stafford, & Stauffer, 2000; van Zyl, Neudecker, & Nel, 2000; Bonett and Wright, 2000; Bonett, 2002; Koning & Franses, 2003; and Duhachek & Iacobucci, 2004 for other contributions to the determination of inferences from sample alphas to population alphas.)

Consider the hypothetical example in Chapter 8 of the alpha of .449 for the four-item test administered to five persons, and suppose that those five persons constitute a random sample from some "infinitely large" population of persons. If you would like to estimate the population alpha from the sample alpha "within bounds"--a 95% confidence interval, say--you would calculate the endpoints of the interval as follows (see Feldt, 1965 for the formulas, which are also provided, along with the data for this same example, in Knapp, 1991):

Lower limit = $1 - F_L\ (1 - \alpha)$, where F_L is the 97.5th percentile of the F sampling distribution for n -1 and (n - 1)(k - 1) degrees of freedom (n is the sample size and k is the number of items), which for our example is equal to 1 - 4.12 (.551) = -1.270 (trust me).

Upper limit = 1 - F_U (1 - α), where F_U is the 2.5th percentile of the F sampling distribution (and everything else is the same as for the lower limit, which for our example is equal to 1 - .114 (.551) = .937 (trust me again).

Therefore, given a sample alpha of .449 for a four-item test and a sample size of five persons, "reasonable limits" for the population alpha are -1.270 and .937. You will note two things about that interval: (1) the lower limit is negative (that also happened in the previous section for the height example) and is even less than -1; and (2) the interval is awfully wide (but that's what you get when you have such a small sample). As is the case for many other statistics, the confidence interval approach can also be used to test hypotheses concerning possible values of the population alpha. For our example, almost any "candidate" for the population alpha would be unrejectable!

Goodman & Kruskal's gamma, Kendall's tau-b and Williams' correlation

In the previous chapter I described three methods for determining the reliability of a measuring instrument that produces ordinal measurements: gamma (Goodman & Kruskal, 1979), tau-b (Kendall & Gibbons, 1990), and a special kind of canonical correlation (Williams, 1952). If you have a sample gamma or a sample tau-b and you want to construct a confidence interval for the corresponding population parameter, Agresti (1984) explains how to do it (it's not easy!). If you've calculated Williams' canonical r for your sample, the general approach to constructing a confidence interval around a sample canonical correlation coefficient would provide the basis for the appropriate inference (also not for the statistically faint of heart, and works only for large sample sizes; see Glynn & Muirhead, 1978).

Cohen's kappa

As far as kappa is concerned, Cohen (1960), Fleiss, Cohen, and Everitt (1969), and Fleiss (1971) have provided approximate formulas for confidence intervals (and thus hypothesis tests). Those formulas are a real mess (the formulas for "uncorrected" proportion agreement are a picnic by comparison--confidence intervals and significance tests for proportions have been around for a long time), so I won't give them to you. (If you're interested, see any of the three sources just cited). For Fleiss's (1971) interesting example of six raters providing ratings for six subjects each (but not the same six--there are missing data, by design) on five diagnostic categories, the sample kappa is equal to .430, its standard error is approximately .028, and the 95% confidence interval for the population kappa is .430 ± 2(.028), i.e., from .374 to .486.

Although there are other kinds of reliability coefficients encountered in the literature, parallel-forms, test-retest, alpha, and kappa coefficients constitute at least 90% of those actually used by practicing researchers, so statistical inferences for those other statistics will not be pursued here. (If you care about

statistical inferences for stepped-up split-half reliability coefficients, for example, see Kristof, 1963a,1964; and Lord, 1974.)

Reliability and power

In Chapter 4 on attenuation I pointed out that correlation coefficients and differences between means are both reduced by the unreliability of measuring instruments, and the emphasis was placed on the former, with an appeal to the correction-for-attenuation formula

$$r_{TxTy} = r_{XY} / (\sqrt{r_{XX}})(\sqrt{r_{YY}})$$

Here I would like to shift the emphasis to the difference between means, not because correlations are any less important but because it is more conventional to talk about power in the context of mean differences.

To refresh your memory of introductory statistics, power is the probability of rejecting a false null hypothesis, which is what most researchers would like to do. (If the null hypothesis, which is usually the "nothing is going on" hypothesis, is not true it should be rejected in favor of an alternative hypothesis that "something is going on".) But what happens if you have a less-than-perfectly-reliable instrument upon which that hypothesis is based? The short answer is that power is reduced and you're less likely to be able to reject a false null. That needs to be shown, and I will now proceed to do so.

Consider the all-pervasive "pooled" t test of the significance of the difference between two independent sample means (experimental and control, male and female, or whatever). The formula for "t for two" is:

$$t = \frac{M_1 - M_2}{\sqrt{ S_p^2 / n_1 + S_p^2 / n_2 }}$$

where M_1 and M_2 are the obtained means for sample 1 and sample 2, n_1 and n_2 are the corresponding sample sizes, and S_p^2 is the pooled variance $(n_1 S_1^2 + n_2 S_2^2)/ (n_1 + n_2 - 2)$ for the corresponding obtained variances S_1^2 and S_2^2 .

Unreliability presents no problem for the numerator of the t ratio, since the obtained means are at least approximately equal to the true means by Theorem #1 of Chapter 3. But those two obtained variances are sums of true variances and error variances, so if the instrument used to get the data is subject to any unreliability the obtained variances are too large (the true variances having been inflated by measurement error), the pooled variance is also too large, t is too small, and you are less likely to be able to reject the null hypothesis, no matter whether it is true or false, so power is lower than it would be if the instrument were perfectly reliable. (Did you follow that?)

If you have two dependent samples, the formula for t is different, but the issue is the same (the obtained variance in the denominator is too large). Likewise for more than two samples, independent or dependent, where the within-sample denominator of the F ratio used in the appropriate analysis of variance is also too large, F is correspondingly too small, and power is lower.

Unreliability of change scores can also decrease the power of a significance test for experiments in which the mean change on a dependent variable for an experimental group is compared with the mean change for a control group where a pretest/posttest design has been employed. For an interesting exchange of opinions concerning this problem, see Overall and Woodward (1975, 1976), Fleiss (1976), and Nicewander and Price (1978).

You wouldn't believe the number of people who have studied the problem of the effect of unreliability on power. For particularly good explanations of the problem and suggestions regarding what to do about it I recommend the articles by Cleary and Linn (1969b), Subkoviak and Levin (1977), Sutcliffe (1980), Zimmerman and Williams (1986), and Bacon (2004); and Chapter 8 of Aiken and West's (1991) multiple regression textbook.

Sample size for reliability studies

Closely related to power is the matter of the appropriate sample size to use in a reliability study. Let us re-consider the above example of testing the "null" hypothesis that the population reliability coefficient is equal to .80 against the alternative hypothesis that the population reliability coefficient is less than .80. There we were "stuck" with a sample of only seven observations and we could not reject the .80 even though our sample reliability coefficient was only .50. The more interesting research question is: What size sample is appropriate for testing those two hypotheses against one another so that if the null is true we would have a reasonably high probability of arriving at that conclusion, and if the alternative is true we would also have a reasonably high probability of arriving at that conclusion? In formal inferential statistical parlance, we would like the probability of making a Type I error and the probability of making a Type II error to both be small.

The Type I error part is easy: Choose a sufficiently stringent alpha level (not to be confused with Cronbach's alpha!) so that the probability of rejecting a true null is small (.05 is conventional and should be fine). The Type II error part is hard. First of all, we can't even get off the ground unless we make the alternative hypothesis as specific as the null; that is, we need to hypothesize another value for the population reliability coefficient that we're willing to believe is true if the null-hypothesized value of .80 is not. Let's say that the non-null hypothesized value is .90 (that our test is more reliable than the typical test--how optimistic can you get?!). Secondly, we need to also choose a sufficiently stringent beta level (not to be confused with the beta weight in regression analysis!), so that the

probability of failing to reject a false null is small. (Do you follow both of those requirements?) Those two probabilities don't have to be equal--the consequences of being wrong may not be the same--so let us choose the beta level to be .20. Since power is equal to 1-beta, we will have implicitly chosen that to be .80 (Cohen's "default" power). We're now all set (the nature of the two hypotheses necessitates a one-tailed test), because there exist formulas and handy tables for determining the optimal sample size for testing the .80 vs. the .90--optimal in the sense that if our sample size is less than that number our power will be less than .80 and if our sample size is greater than that number we will be incurring unnecessary costs.

I recommend that you see Cohen (1988) for sample size determination in general. And see Donner and Eliasziw (1987); Eliasziw and Donner (1987); Walter, Eliasziw, and Donner (1998); and Feldt and Ankenmann (1998, 1999) for sample size determination for reliability studies in particular. For example, if you wanted to test the "null" hypothesis that the reliability coefficient in the population is equal to .80 (what previous investigators have gotten, say) against the more promising alternative hypothesis that it is equal to .90 (your claim for your instrument), using the .05 significance level and with a desired power of .80, the required sample size is approximately 46.

The effect of (un)reliability on confidence intervals in general

As I have pointed out several times already, interval estimation generally subsumes hypothesis testing, so it should come as no surprise to you that if a statistic has been determined by using one or more less-than-perfectly-reliable measuring instruments, the confidence interval for the corresponding parameter will be necessarily wider than it would be for perfectly reliable instruments. This is most easily seen by considering the familiar formula for the confidence interval for most statistics, i.e., statistic ± margin of error. The margin of error is typically some multiple of a standard error of the statistic. The standard error in turn is a function of an obtained standard deviation, which is the square root of an obtained variance, which is too large since the true variance is inflated by error variance (see Theorem #2 in Chapter 3) whenever the reliability coefficient is less than 1.

The determination of sample size for confidence intervals of tolerable width is similarly affected. The more unreliable the instrument(s), the larger the sample size necessary to provide the desired "coverage", all other things (such as the specified confidence coefficient) being equal.

"Bootstrapping" reliability coefficients

Dunn (2004) provides a particularly interesting example of the use of the non-parametric bootstrap for sample-to-population inferences concerning reliability statistics, when one is unable or unwilling to make certain assumptions about the

underlying population distribution and/or the theoretical sampling distribution of a particular indicator of reliability is unknown or mathematically intractable.

Our flow meter example (re-revisited)

I would like to close this chapter by returning to our expiratory flow example and showing how a number of statistical inferences might be made for such data, if the data were treated as coming from a sample rather than constituting an entire population. You may recall that 17 subjects were measured twice with each of two versions of an instrument, the standard Wright meter and the Mini Wright meter. One of the interesting results was a correlation of .983 between the first and the second measurements obtained with the standard meter. If those 17 people had constituted a simple random sample from a larger population of interest (they were actually a handy, "convenience" sample), the 95% confidence interval for the population correlation (reliability coefficient) would be determined as follows:

r = .983 (Fisher's z equivalent = 2.400)
n = 17
standard error of Fisher's z = $1/\sqrt{(n-3)}$
= .267

The 95% confidence interval for Fisher's z in the population is 2.400 ± 2 (.267), i.e., 1.866 to 2.934. The 95% confidence interval for ρ (the population reliability coefficient for the Wright meter) is from approximately .95 to .99, a surprisingly tight confidence interval for such a small sample size (a ceiling effect?), and indicative of very high within-meter reliability.

Another interesting correlation is the correlation between the First Y for the Mini Wright meter and the Second Y for the Mini Wright meter, .967. Estimating the population correlation ρ for that meter would proceed in the same fashion and would produce a similarly "high and tight" interval.

But perhaps the most interesting relationship is that between the Wright and the Mini Wright. Are the two instruments compatible? Lord (1957b, 1973) provided a test of the hypothesis that two instruments measure the same thing (in this case, expiratory flow), except for possibly different errors of measurement, different metrics (e.g., inches vs. centimeters), and/or origins of measurement (e.g., the arbitrary zero points for Fahrenheit degrees vs. Centigrade degrees). The calculations are a bit complex (see Lord's articles for the actual formulas), but as you might expect from the discussion of the reliability of differences in Chapter 6 they involve all six within-instrument and between-instrument correlations and all four standard deviations (or functions of them), as well as all four means. Anyhow, for this example it turns out that the hypothesis of compatibility is rejected; i.e., the two instruments appear to be measuring

different things. (See my comments regarding Subjects 6, 7, and 16 when this example was first introduced in Chapter 3.)

Random samples vs. "convenience" samples

All of the above discussion and the various formulas, strictly speaking, apply only to simple random samples, i.e., samples drawn from populations in such a way that every combination of n objects in the population has an equal and independent chance of being included in the sample. Such samples are exceedingly rare in actual research. Some (perhaps most) researchers claim that inferential statistics are also appropriate for non-random samples that are conveniently obtained rather than based upon any chance process. The matter is very controversial. "Liberal" researchers "regard" their samples as random samples from hypothetical populations "like these"; or (b) they use inferential statistics in order to provide an objective basis for determining whether or not to get excited about a particular sample result; or both. "Conservative" researchers reject both of those arguments, because regarding and having are two different things, they have no interest in hypothetical populations, many aspects of inferential statistics are subjective, not objective (e.g., the choice of significance level), and theory, not statistical inference, should be the basis for "excitement". I'm with the conservatives; inferential statistics are vastly overused (in my opinion).

Additional reading

For an Excel program that computes confidence intervals for various reliability coefficients, see Barnette (2005). For general information regarding statistical inferences for reliability indices, see Sutcliffe (1958), Kristof (1963b; 1970), and Charter (1999).

CHAPTER 12: A very nice real-data example

In the previous chapters of this book I have made extensive use of two examples: (1) a set of hypothetical data for the heights of seven basketball players; and (2) a set of real data for 17 subjects measured twice with each of two peak expiratory flow meters. Although both of those examples were helpful (I hope) for illustrating various principles of classical reliability theory, neither is representative of the kinds of measurement situations that researchers typically encounter. In this chapter I would like to apply many of the concepts that we have discussed to a much larger set of real data. The data served as the empirical basis for an article entitled "Is self-reported height or arm span a more accurate alternative measure of height?" (Brown, Feng, & Knapp, 2002) and they can be found in Appendix A, by courtesy of the senior author of that article, Dr. Jean K. Brown, University at Buffalo, State University of New York.

Background and the study itself

The measurement of height is essential for the calculation of body surface area and body mass index, which are often the basis for healthcare practices such as the determination of drug dosages and weight-reduction therapies. The stadiometers that are commonly found in doctors' and nurses' offices are very useful instruments for measuring people's heights. There is one practical problem, however: The measuree must be able to stand up straight. That is difficult, if not impossible, for some people, e.g., hospitalized patients who are suffering from multiple sclerosis. One suggested alternative way to "measure" height is to ask the person what (her)his height is and proceed from there. But are self-reported heights reliable? A few researchers have studied the problem of the reliability of self-reported heights and/or the validity of self-reported heights (e.g., Pirie, Jacobs, et al., 1981; Stewart, 1982; Larson, 2000). Another suggested alternative is to use arm span as a surrogate measure for height (e.g., Engstrom, Roche, & Mukherjee, 1981; Steele & Mattox, 1987; Kwok & Whitelaw, 1991; Parker, Dillard, & Phillips, 1996), but the reliability of arm span measurement is also questionable. Following upon some earlier work on arm span alone (Brown, Whittemore, & Knapp, 2000), Brown, Feng, and Knapp (2002) undertook a study of 409 subjects who self-reported their heights and had their arm spans measured (twice each) and had their heights measured (also twice each). What makes that study so interesting and unique is that the measurements were taken by 82 different measurers (an average of about five subjects each) with 82 different metal rules, which had been checked against a Stanley™ model 33-158 rule. (See Brown, Feng, & Knapp, 2002, for the procedural details.) The study's primary purpose was to compare the criterion-related validity (see Appendix B) of self-reported height with the criterion-related validity of arm span measurement (the criterion being actual measured height). It is the reliability data (two measurements of height and two measurements of arm span), however, that are of principal concern here.

Over-all parallelism

The mean of the first set of height measurements is 66.84 and the variance is 15.29. The mean of the second set of height measurements is 66.87 and the variance is 15.37. Because the means aren't identical to one another, nor are the variances, the two sets of height measurements aren't perfectly parallel, but they're certainly "close enough for government work"!

For the arm span measurements, the mean of the first set is 67.93, with variance 20.25; and the mean of the second set is 67.97, with variance 20.43. Those two sets of measurements are also not perfectly parallel, but very close thereto.

[For the validity portion of their study, Brown, Feng, & Knapp, 2002 quite properly chose to use only the first height measurement as the dependent variable, and the first arm span measurement as one of the independent variables, on the off chance that the measurers might have been influenced by their knowledge of the first measurement when making and recording the second measurement.]

Over-all reliability

The correlation between the first and the second height measurements (an estimate of the reliability coefficient for the measured heights) is .997, an indication of extremely high reliability. The correlation between the first and the second arm span measurements (an estimate of the reliability coefficient for the measured arm spans) is, interestingly, also .997, and also indicative of excellent reliability. But both of those reliability estimates are a bit deceiving. The mean absolute difference (the statistic emphasized by Brown, Feng, & Knapp, 2002) is .17 inches for the heights and .21 inches for the arm spans. Those look like small numbers (less than a quarter of an inch). What they convey is that if we use those metal rules we're likely to be "off" from true height or true arm span by somewhat less than 1/4 inch on the average. But that's "on the average". We could be "off" by a lot less (that's the good news) or by a lot more (that's the bad news). If you look at the actual raw data in Appendix A you will see that there were several differences of 0 inches between the two height measurements but there were also a few fairly large discrepancies, e.g., for IDs 148, 163, and 173--differences of an inch or more, and all three with the second height greater than the first height (they stretched a little between measurements?). The same sort of thing happened for arm span (several zero differences; a few large differences, e.g., for IDs 192, 217, and 247).

The 82 measurers

Now for the fun part. Recall that there were 82 different measurers using 82 different steel rules to measure the heights and the arm spans of their respective subjects. That raises several interesting questions:

(1) Are those steel rules equally dependable? Some evidence for the answer to this question was provided in the "calibration" phase of the study by the distribution of the measurements taken of a cupboard, a human arm span model, and two human height models. Data were available for 80 of the 82 measurers who measured the cupboard ("true" width = 48 3/16 or 48.1875 inches--the measurers were asked to report their measurements to the nearest sixteenth of an inch). Their actual measurements ranged from a low of 47.50 to a high of 48.3875 (mean = 48.153 and variance = .01568). Data were also available for 79 measurers of the arm span model ("true" arm span = 66.4375 inches, range from 65.875 to 66. 875, mean = 66.356 and variance = .04397); for 43 measurers of one height model ("true" height = 62.9375 inches, range from 62.1875 to 63.50, mean = 62.631 and variance = .04726); and for 34 measurers of the other height model ("true" height = 67.00 inches, range from 66.25 to 67.25, mean = 66.695 and variance = .06970). Those numbers are very interesting, albeit not directly comparable since they're not all based on the same measurers. For instance, (a) none of the means of the obtained measurements are equal to the Platonic true values (that is particularly bothersome for the two height models, where there is a decided downward bias); (b) the agreement was much greater for the cupboard (variance = .01568) than for the human models (variances = .04397, .04726, and .06970); and (c) the agreement was better for the shorter height model (variance = .04726) than for the taller height model (variance = .06970). So I guess the answer to the question (Are the 82 rules equally dependable?) is "no".

(2) Were the steel rules parallel in the main study? We have established the parallelism of the two OCCASIONS but haven't yet faced up to the parallelism of the 82 instruments/measurers. Instrument and measurer are confounded with one another here, because no measurer used more than one instrument, so if there is any unreliability we cannot determine whether it is "the instrument's fault" or "the measurer's fault". (Do you follow that?)

(3) (a corollary to the previous question) Are the data "poolable" across the 82 instruments? If, for example, one instrument yielded first and second obtained heights of 65 and 65 for Person A, and 70 and 70 for Person B, respectively; but another instrument yielded first and second obtained heights of 65 and 70 for Person C, and 70 and 65 for Person D, respectively; the data for those two instruments would not be poolable because there is a perfect positive relationship (Pearson r = +1) between the height measurements for the first instrument and a perfect negative relationship (Pearson r = -1) for the second instrument.

(4) (the key question) If the instruments are not parallel, and the data are not poolable, what do we do about it?

In the last column of the data in Appendix A there is a code number indicating who measured whom. Although there is a formal test of the parallelism for several forms (see Gulliksen, 1950 and Lord, 1964), it's a bit of a mess. For our purposes here I will briefly address the matter of parallelism, try to determine "the best measurer" and "the worst measurer", and make a judgment regarding the "poolability" of the data across measurers. By "the best measurer" I mean the one for whose data there is the highest reliability coefficient and the lowest mean absolute difference, and by "the worst measurer" I mean the one for whose data there is the lowest reliability coefficient and the highest mean absolute difference.

The assessment of instrument parallelism is a bit tricky when there is more than one measurer but the measurers have measured different measurees. For the Brown, et al. study you can't compare the obtained means and standard deviations for the 82 height and arm span measurers as you would several forms of a test administered to the same persons, because the true score distributions could be vastly different if, for example, one measurer had measured some very tall persons and another measurer had measured some very short persons, which in fact appears to be the case here. (The mean--obtained and true by an appeal to Theorem #1--first heights range from 62.20 for Measurer 79 to 71.37 for Measurer 21, and the mean first arm spans range from 63.60 again for Measurer 79 to 72.30 for Measurer 37.) But you can compare the first and second measurements of each characteristic (height and arm span) within each measurer. An inspection of the data in Appendix A reveals that four of the measurers (# 37, 57, 60, and 71) had identical values for their first and second height measurements AND for their first and second arm span measurements, so their “forms” were perfectly parallel. On the other hand, one measurer (#44) had rather discrepant values for first and second heights (mean = 65.08 and variance = 2.83 for first height; mean = 64.40 and variance = 2.34 for second height) and another measurer (#18) had similar discrepancies for first and second arm spans (mean = 66.70 and variance = 3.29 for first arm span; mean = 66.25 and variance = 2.95 for second arm span).

[Caution: It is conceivable, but hopefully not the case, that Measurers 37, 57, 60, and 71 "dry-labbed" their data by taking only first measurements and making their second measurements conform to their first measurements. The one measurer whose measurements are perhaps the most suspect is Measurer 57 who reported all measurement to only the nearest half-inch and had no discrepancies whatsoever! The reliability coefficient for the other three of those aforementioned measurers (#37, 60, and 71) is also 1 and the mean absolute deviation is 0 for both height and arm span, so there could be some "hanky-panky" going on there also. Some students involved in research studies have been known to try to "help out" the principal investigators by providing them with

data they think the investigators want; others have been known to try to "louse up" the investigators.]

I nominate Measurer 44 (and/or his(her) metal rule) for the "worst" measurer, with reliability coefficients of .937 for height and .928 for arm span, accompanied by mean absolute differences of .776 and .698, respectively. Do you agree?

Parallelism of measurers can be studied by comparing the individual measurer reliability coefficients and the individual mean absolute differences. I have already cited the ranges, from the lowest reliability coefficient of .937 to the highest reliability coefficient of 1 for height and from the lowest reliability coefficient of .928 to the highest reliability coefficient of 1 for arm span. The correspondingly lowest mean absolute difference for height is 0 and the highest is .776; the correspondingly lowest mean absolute difference for arm span is also 0 and the highest is .698. It's a judgment call as to whether or not that range of values is indicative of non-parallelism.

That brings us to the matter of the "poolability" of the data. Clearly, all of the measurers did not produce equal means, variances, reliability coefficients, and mean absolute differences. My personal judgment is that they are not terribly discrepant, but serious consideration could be given to deleting that "worst" measurer (the next worst mean absolute difference for height is .500 and the next worst mean absolute difference for arm span is .600, for example, and not for the same measurer) and any of those "best" measurers whose data might be suspect.

Tidbits

A close examination of the data in Appendix A reveals that the decimal portions of some of the height and arm span measurements are not decimal equivalents of sixteenths of an inch. (All of the measuring instruments were graduated in eighths of an inch, with the capability of interpolation between adjacent values.) For example, Subject 7's first arm span measurement, as reported by Measurer 2, is 63.20 inches, and the .20, no matter whence it was rounded, is not a multiple of .0625 (which is the decimal equivalent of 1/16). Was the measurement that was actually made 63.1875 (= 63 3/16), 63.25 (= 63 1/4), or just what? Unfortunately we'll never know. As I indicated above, the 82 measurers were "on their own" after their original calibration exercises. Dr. Brown did not, and could not, monitor each one of them while they were measuring their family members or friends. This is the sort of thing that happens in real-world research, and yes, it does contribute to the unreliability of the measurements. [Dr. Brown told me that she had a few foreign students in her class who were accustomed to the metric system (meters, centimeters, millimeters, etc.), rather than the British/American system (feet, inches, fractions of inches, etc.), and although she cannot, and should not, identify who Measurer 2 was, it is conceivable that (s)he read off the two "tick marks" to the right of the

63-inch mark as two tenths of an inch instead of two eighths of an inch. (The decimal portion of her (his) second arm span measurement for Subject 7 was also not a multiple of 1/16; but the decimal portions of both height measurements were. Go figure!)]

Although self-reported height is not a variable of direct concern here (it was of vital concern to Brown, Feng, & Knapp, 2002), it is interesting to note that most people appear to have self-reported their heights to the nearest inch (they were not told what precision to use), but there were at least 63 people who did try to estimate their heights more precisely (the 63 subjects who had a self-reported height that did not end in .00). If their first actual height measurement can be defended as their "gold standard" height, more of them over-estimated than under-estimated.

CHAPTER 13: Special topics

In this chapter I would like to discuss several matters that are of less importance (in my opinion) than those that we have already covered and/or didn't fit in very well in any of the previous chapters.

Some other conceptualizations of reliability

The first of these topics is concerned with the competing approaches to classical reliability theory of generalizability theory, item response theory, and structural equation modeling as the foundations for measurement in the social sciences. (Once again, those of you who are primarily interested in physical science instruments might want to skip this section.) I will admit at the outset that I am far from an authority on any of these alternatives, but I know enough about them (I hope!) that I can compare and contrast them with what has been said so far in this book, and to point out their advantages and disadvantages relative to the classical approach.

Generalizability theory

Generalizability theory is often said to have been originally proposed in an article by Cronbach, Rajaratnam, and Gleser (1963) and later in book form by Cronbach, Gleser, Nanda, and Rajaratnam (1972), but its history actually pre-dates both of those sources (see Brennan, 1997). It has been subsequently summarized in somewhat more readable form by Brennan (1998, 2000, 2001a, 2001b), Shavelson, Webb, and Rowley (1989), Shavelson and Webb (1991), Dunn (2004), and others. It differs from classical reliability theory in one major respect: It attempts to "break down" the measurement error component of an obtained score into several sources, rather than lump them all together into one big E. Generalizability theorists want to know how much of the variation in obtained scores is attributable to forms (when you have more than one form), how much of the variation is attributable to items (when there are items), how much to occasions (when people are tested more than once), etc. As a result they lean heavily on the analysis of variance in so doing. And because one can identify various kinds of errors, there is a reliability coefficient (called, naturally enough, a "generalizability coefficient") associated with each kind of error. There are also (hang on to your hats) relative generalizability coefficients and absolute generalizability coefficients (see Burns, 1998, for an interesting example).

Item response theory

Item response theory (IRT), formulated by Thurstone (1925), Lord (1952), Rasch (1960), and others (see Bock, 1997, Wright, 1997, and McDonald, 1999 for accounts of its history), and popularized by Hambleton, Swaminathan, and Rogers (1991), Hambleton and Jones (1993), Hambleton and Slater (1997), and

others, starts with the premise that every test item j has an "item characteristic curve" such that for each examinee i there is some (unknown but estimable) probability p_{ij} that (s)he will answer the item correctly (or, for the case of affective measurement, provide the favored response), and a good measuring instrument should be constructed in such a way that those item characteristic curves have certain desirable properties. The mathematics gets pretty heavy (with lots of emphasis on exponents and logarithms), and there are indices of measurement error that correspond to the reliability coefficients, standard errors of measurement, etc. of classical reliability theory (see, for example, Lord, 1980; Wright & Masters, 1982). In Rasch measurement there are two kinds of reliability coefficients, one for the reliability of item separation (the proportion of between-item variance that is not associated with measurement error) and one for the reliability of person separation (the proportion of between-person variance that is not associated with measurement error). The interested reader is referred to the Wright and Masters monograph for the respective formulas and their interpretations.

In his very long and technical article, Mislevy (1996) attempted to reconceptualize test theory in order to provide a more defensible basis for measurement in cognitive and developmental psychology. Holland & Hoskens (2003) discussed classical reliability theory as a general version of item response theory. And Doran (2005) questioned whether the information function in item response theory is an indicator of reliability. [My answer is "no" (I don't think it has anything to do with consistency), but if you're knowledgeable about IRT please read Doran's article and see what you think.]

Structural equation modeling

Structural equation modeling (SEM), also known as the analysis of covariance structures (and one or two other similar designations), which originated with the work of Karl Joreskog (see, for example, Joreskog & Sorbom, 1979), is a currently-popular method for investigating the causal connections between underlying scientific "constructs" or "latent variables" (what I called "attributes" in Chapter 2) as well as their various empirical operationalizations ("indicators" or "manifest variables"). It can be thought of as a generalization of path analysis, which is itself a special type of regression analysis. There are two parts to a structural equation model: the measurement model, in which the hypothesized relationships between the constructs and the indicators are tested; and the structural model, in which the relationships between the constructs are investigated. (See Bollen, 1989; Mueller, 1996; and Drewes, 2000 for clear discussions of the basic concepts in structural equation modeling. Bollen's pages 206-223 are particularly good for clarifying the difference between classical reliability and SEM reliability, and the difference between SEM reliability and SEM validity.) SEM advocates define the reliability of an indicator variable as the square of its "loading" on the construct it is alleged to tap. (The constructs are also often referred to as "factors"--confirmatory factor analysis is an integral

part of SEM.) Although the theoretical constructs can be assumed to be measurement-error-free, if that assumption is not warranted their reliability can also be estimated. Fornell and Larcker (1981), Miller (1995), Raykov (1997), and Hancock and Mueller (2000) have all derived measures of a construct's reliability as a function of the reliabilities of its indicators. The Hancock and Mueller statistic, which they call H, appears to have the most desirable properties, e.g. that the reliability of the construct can never be less than that of the indicator of highest reliability.

Norm-referenced vs. criterion-referenced reliability

There is a reasonably large literature devoted to the way instrument reliability is handled within a "norm-referenced" framework and how it is handled within a "criterion-referenced" framework. (We have been operating in an "un-referenced" or "raw" framework.) The terms "norm-referenced" and "criterion-referenced" come from educational measurement. The former is concerned with the interpretation of a measurement with respect to other measurements that have been obtained with the same instrument, a percentile rank being the most familiar type. (Example: "Mary's obtained score of 37 on a spelling test put her at the 91st percentile when compared to a national sample of third-graders.") The latter (criterion-referenced, sometimes called "domain-referenced") is concerned with either or both of two things: (1) how much of the domain has been "bitten off"?; and (2) is that a "passing" performance? (Example: "John spelled 82% of the words correctly, which was below the "cutting point" for progressing to the next lesson.")

Educators whose instruments are norm-referenced have adopted reliability theory pretty much "as is", with its emphasis on correlation coefficients that are indicative of relative associations between variables. Those whose instruments are criterion-referenced have tended to favor a modified version of reliability where the emphasis is on measurement error in the vicinity of the cutting point. What is of even greater interest is the reliability of the pass-or-fail decision. In parallel-form situations, for example, the issue of whether a person passes both Form A and Form B or fails both Form A and Form B takes precedence over how high the correlation is between the two forms.

For more on criterion-referenced reliability see Livingston (1972, 1973), Harris (1973), Swaminathan, Hambleton, and Algina (1974), Huynh (1976), Subkoviak (1976, 1978), Wilcox (1978), Livingston and Wingersky (1979), Huynh and Saunders (1980), Peng and Subkoviak (1980), Berk (1980), Traub and Rowley (1980) [1980 was a good year for criterion-referenced reliability!], and especially Traub (1994), Feldt (1996), Chase (1996), and Puhan and Gall (2005). Chase's article actually discusses a method for estimating the reliability of a criterion-referenced instrument before it's ever administered!

(Non-)Independence of Observations and Unit-of-analysis problems

All four measurement theories (classical reliability theory, generalizability theory, item response theory, and structural equation modeling) typically assume that the unit of analysis (the object that is measured) is an individual person and not some aggregate of objects such as a family, a classroom, or a hospital. And if the individuals are "nested" in such aggregates the nesting must be taken into account in order to cope with the problem of the non-independence of observations within aggregate in conjunction with the greater independence of observations across aggregates.

A not uncommon occurrence is the ability to be able to measure only at one of those "higher" levels of aggregation. Consider, for example, an instrument that has been designed to measure the coherence of a family (how "together" it is). The object of measurement is the family itself, and the measuree is usually one member of each family who is asked to perceive family coherence; or, if more than one family member is measured, the data are immediately aggregated to the family level (usually by taking the average of the individual measurements) because non-independent observations would result if more than one measurement for a given family were to be used. No matter which way the data are obtained, the problem is that evidence regarding the reliability (and validity) of the measuring instrument has probably been gathered with the individual person, not the family, as the unit of analysis. A high (or low) reliability coefficient for individual observations does not necesarily imply a high (or low) reliability coefficient for aggregate observations.

There is an additional problem whenever this "nesting" arises (students nested within classrooms, classrooms nested within schools, etc.), even when the unit of analysis is the individual object, and that is the matter of whether some statistic such as a correlation coefficient should be calculated within each of the aggregates, across the aggregates, or both. In the previous chapter, for example, there were 82 sets of aggregate data by virtue of the fact that there were 82 measurers nested within the entire data set. We found that the correlation between first actual measured height and second actual measured height was .997 (n = 409) across all of the measurers, but the within-measurer correlations ranged from .937 (n = 5) to 1 (n = 5).

One of the first methodologists who tried to cope with the unit-of-analysis problem in a measurement context was Sirotnik (1980), who used as illustrative examples situations involving the measurement of the organizational climate of schools (a matter that is substantively similar to the coherence of families). Drawing on the unit-of-analysis literature that was available at the time (there is now a much larger literature, and it goes by the fancier name "hierarchical linear modeling"--see, for example, Raudenbush & Bryk, 2002), he showed that the usual formulas for estimating measurement error for individual objects must be modified when applied to aggregate data.

Following later on Sirotnik's work, Verran, Mark, and Lamb (1992) summarized the problem for nursing researchers who have aggregate data and who are concerned with the effect of unreliability on such data. And Forbes and Taunton (1994) provided an application to data aggregated to the organizational level. (See also Franzen & Derryberry, 1932b and Brennan, 1975 for harbingers of what was to come in Sirotnik's work.) More recently, Waller (2008) argued that "commingled" samples, i.e., two or more samples pooled together and treated as a single sample, can have serious effects on the reliability of the instrument(s) used for those samples.

As I indicated in Chapter 9, the intraclass correlation coefficient plays a key role in unit-of-analysis problems, but in that context it is usually referred to as a correlation ratio, given the symbol η^2, and calculated by dividing the between-aggregate sum of squared deviations from the grand mean by the total sum of squared deviations from the grand mean. For more on the unit of analysis and the independence of observations, see my articles: Knapp (1977, 1984).

Weighting

When an instrument is composed of two or more parts, e.g., test items, and the measurements for the component parts are to be combined in order to produce a total obtained score, there is always the question of how to combine them. Most of the time a simple sum is taken, i.e., the parts are given equal weight, but there are times when it might be better to give them different weights. I have mentioned differential weighting a couple of times in previous chapters (when getting an over-all estimate of the reliability of an item by combining the "reliability for rights" and the "reliability for wrongs", for example--see Chapter 7). Many measurement experts, from Pothoff and Barnett (1932) to Drewes (2000), have been concerned with the problem of how to weight the components so as to produce a weighted sum that has optimal properties, usually maximum reliability. (See Thomson, 1940; Mosier, 1943; Green, 1950; Aiken, 1966, 1988; Stanley & Wang, 1970; Conger, 1974, 1980; Li, Rosenthal, & Rubin, 1996; Li, 1997; Li & Wainer, 1997; Cliff & Caruso, 1988; and tenBerge, 2000.)

An example of the epitome of reliability optimization was given by Li, et al. (1996). They showed that if a test is to consist of two kinds of items (essay and multiple-choice), if each essay item has a reliability of .60 and costs $1.00 to score, if each multiple-choice item has a reliability of .10 and costs $.01 to score, and if the fixed cost (budget constraint) for scoring the test is $7.00, then a maximum reliability (of .97) would be achieved by having the test consist of 200 multiple-choice items and 5 essay items. (See also Wainer & Thissen, 1993 and Kane & Case, 2004 regarding weighted composite scores.)

The two articles by Aiken, 22 years apart (1966 and 1988) are an interesting pair. In the 1966 article he showed that under usual testing conditions (e.g., positive inter-item correlations) equal item weighting and differential weighting tend to put

examinees in essentially the same rank order, the implication being that differential weighting is rarely worth the effort. But in the 1988 article he provided a computer program for determining the maximum reliability of a weighted composite!

Other researchers have investigated other weighting problems. Ebel (1965) and Rippey (1968, 1970), for example, studied the effect of "confidence weighting" on the reliability of a multiple-choice tests. Here weights are assigned to each choice for each item by the examinees, indicating how certain they are of the correctness of that choice, and/or to each item, indicating how certain they are of the correctness of their response to the item. (See also the previous work by Davis & Fifer, 1959.) The results were mixed; sometimes the test versions asking the examinees for confidence weighting were found to be more reliable than the traditional test versions where no such information is asked, but sometimes they were found to be less reliable. (It's amazing what things researchers choose to study, isn't it?)

Missing-data problems

No problem is more frustrating to substantive researchers than that of missing data. They go to great lengths to design their instruments with careful instructions concerning how measurements are to be taken, only to later discover that one or more measurees refused to provide certain information, the measurers didn't record the information properly, the people entering the data for analysis purposes omitted the information, or some other goof-up. What effect does that have on the estimation of the reliability of the instrument? It depends.

If the "missingness" is built into the study (see, for example, Fleiss's 1971 article on statistical inferences for Cohen's kappa that was cited in Chapter 11), the only sacrifice is a loss of some sensitivity in the data. But if the "missingness" happens a lot and in unplanned places, e.g., every person has missing data for different items, the effect can be very severe. The only remedies are imputation and deletion. Imputation strategies involve the estimation of what each datum "might have been"; deletion strategies involve the elimination of some of the non-missing data (e.g., dropping all of a person's data if (s)he has 5% or more missing data), and that approach therefore results in even more missing data! The principal objective of an imputation strategy is to salvage all of the non-missing data while at the same time creating a full data set that has desirable properties. The principal objective of a deletion strategy is to make life easier (e.g., to be able to calculate a matrix of correlation coefficients that are always possible for real data). There are a number of different imputation strategies available (e.g., mean substitution, use of the Expectation Maximization algorithm) as well as a number of deletion strategies (e.g., listwise deletion, pairwise deletion). I discuss some of them briefly in my nursing research textbook (Knapp, 1998), but if you're really interested in how to handle missing data the

principal resource is the book by Little and Rubin (2002). [See also Huynh (1986b) and Enders (2003; 2004).]

Some miscellaneous educational testing examples

There have been a number of other interesting contributions to the literature on the reliability of measuring instruments. One of my favorites is the article by Glass and Wiley (1964) in which they showed that the reliability of multiple-choice test scores that are not corrected for guessing (i.e., are scored as R, where R is the number of correct answers) is higher than the reliability of such tests when the scores are corrected for guessing (i.e., are scored as R - W/(c-1), where W is the number of wrong answers and c is the number of choices per item). Plumlee (1952, 1954), Mattson (1965), Zimmerman and Williams (1965), and Traub and Hambleton (1972) also studied the effect of the guessing correction and the degree of speededness of a test on reliability and/or validity. Another of my favorites is the article by Ebel (1969b) in which he showed that the finer the scale for grading test performances, the higher the reliability. That shouldn't be surprising to you after you've studied reliability theory (the basis of Ebel's proof), but it is fairly common folk "wisdom" that for an essay test, for example, you can't make fine distinctions between examinees so you should use a very small number of grade categories. (The extremists recommend just two: pass and fail.) Ebel includes a table that shows, among other things, that reducing the number of categories from five to two results in a loss of reliability from .85 to .63 for an instrument whose maximum possible reliability is .95. In addition to a loss of reliability, the additional cost of reducing the number of categories for any non-continuous variable is a loss of predictability and of power, as Cohen (1983) so clearly pointed out about 20 years ago and MacCallum, Zhang, et al. (2002) re-emphasized. Later, Feldt (2005) provided a method for deducing the reliability of dichotomized and trichotomized scores from the reliability of a continuous score that is “broken down” into two or three categories.

But perhaps my most favorite of all is one of the oldest of all, an article written by Ruch and Stoddard (1925) on the same topics with which Lord (1944), Glass and Wiley (1964), Traub and Hambleton (1972), and Ebel (1969a) were concerned, but pre-dating them by several years. In a controlled experiment involving 562 high school seniors in Iowa, they investigated the relative reliabilities of five different types of objective tests: recall, five-choice, three-choice, two-choice, and true-false, with two 50-item parallel forms of each type. (There is a subtle difference between two-choice and true-false, as illustrated by the versions of one of their items: "The American Revolution began in 1762 1775" vs. "The American Revolution began in 1775 True False".) They found (anticipating Lord, Glass & Wiley, Traub & Hambleton, and Ebel?) that as the probability of chance success by guessing increased, i.e., as the number of choices decreased, the split-halves reliability of the total-test scores generally decreased, except for the three-choice version, which was lower than two-choice but higher

than true-false: .811 for recall; .796 for five-response; .598 for three-choice; .737 for two-response; and .555 for true-false. (Their empirical findings agree rather closely, but not perfectly, with theoretical expectations later derived by Lord, 1944. If the assumptions of classical reliability theory are satisfied, reliability increases as the number of choices per item increases and as the number of items increases--see Ebel, 1972 for more on the latter.) Few, if any, educational studies are carried out today with such care and technical expertise. They even estimated what the various reliability coefficients would have been if all students had taken 18.7 minutes to complete 100 items! (18.7 minutes was the median time for the recall version.) The Ruch & Stoddard study is not cited by Glass and Wiley, Traub and Hambleton, or Ebel, but in their partial defense it was an empirical investigation and they were probably interested only in theoretical precedents.

Some more esoteric contributions

Mathematicians are prone to wanting to generalize results or to see where something works and where something doesn't. If they find, for example, that a theorem holds for some k (=2, say), they'd like to know if it holds for all k. And when they present formulas for quotients they always warn the potential user to beware of situations where the denominator might be equal to zero, in which case things "blow up".

Much of the research concerning the reliability of measuring instruments has proceeded in the same way. We've already talked about split-half reliability, where a test is split into two equal parts and the reliability of the total-test is estimated by correlating the parts and "stepping up" that correlation by the Spearman-Brown formula. We've also talked about its generalization to the estimation of the reliability of a test that is k times as long as the one in hand. But what if the test were split into two unequal parts? Or what if the test were split into three parts? Would you believe that both of those problems have already been solved?! See, for example, Feldt (1975) for two unequal parts, and Kristof (1974) for "split-thirds". Kristof (1971), Gilmer and Feldt (1983), and Liou (1989) even worked on the problem of estimating the reliability of a total test when the parts are of unknown length!

CHAPTER 14: The reliability of claims

All of the preceding chapters were concerned with the reliability of actual measuring instruments, where "measuring instruments" were things like math tests, yardsticks, thermometers, etc. In this chapter I would like to extend the concept of reliability to various claims that are often encountered in daily life.

Let me begin with an example: Is global warming a serious problem? Some scientists claim that it is; others claim that it isn't. How might we determine the extent to which such claims are reliable? (The extent of their validity is much more difficult to determine; but as I have tried many times to remind you: this book is concerned almost exclusively with reliability.)

Although it might be too much of a stretch, let us consider the claim of each claimant as an "item" on a "test". The test has two forms. Form A is the "yes" form; Form B is the "no" form. Are the forms "parallel"? We could select a random sample of k scientists from the "yes" population and have their claims constitute Form A. We could correspondingly select a random sample of k scientists from the "no" population and have their claims constitute Form B. The two forms could then be deemed to be randomly parallel, albeit on opposite sides of the issue. Alternatively, we could simply assemble one set of "yes" claims into Form A and another set of "no" claims into Form B and make a subjective determination of their parallelism, as Truman Kelley did when he argued that parallelism is a judgment call (see Chapter 3).

Are the items within form internally consistent? That should be relatively easy to determine, by comparing Scientist i's reasoning with Scientist j's reasoning (i = 1, 2, ...,k) within each of the two forms.

How about the "correlation" between Form A and Form B, which serves so well as an estimate of the reliability coefficient for measuring instruments? Aye, there's the rub, for two reasons: (1) the items in each of the forms are randomly (or purposively) drawn from two different populations (the "yes" population and the "no" population); and (2) what would a high correlation suggest, on the one hand; and what would a low correlation suggest, on the other hand. My overly-optimistic way of coping with both of those problems is to treat the "data" (the items in the two forms) just like points that are made in debates, by considering the cogency of the items as a whole in Form A in comparison with the cogency of the items as a whole in Form B (not their correctness---that would be validity, not reliability). If the two forms contain items of equal or near equal cogency they are reliable. If they don't, they're not. Does that make sense?

Let me conclude this brief chapter with another example. (There is one more example in Appendix F.) Here's the example: Was O.J. Simpson guilty or innocent of the murder of his wife and her friend Ronald Goldman? The

arguments of the prosecutor(s) constitute Form A; the arguments of the defense attorney(s) constitute Form B. I rest my case. (Not really.)

APPENDIX A: The very nice data set

id	age	sex	race	ht1	ht2	arm1	arm2	self	meas
1	54	1	0	61.88	61.50	62.25	62.38	63.00	1
2	52	0	0	64.00	64.13	64.25	64.50	66.00	1
3	23	1	0	64.75	65.00	67.38	67.50	67.00	1
4	23	0	0	68.25	68.00	70.00	70.13	68.00	1
5	22	0	0	68.25	68.50	71.38	71.50	70.00	1
6	50	1	0	62.06	62.19	61.63	61.75	62.00	2
7	23	1	0	63.63	63.75	63.20	63.40	63.00	2
8	23	0	0	75.81	75.94	74.75	75.00	76.00	2
9	22	1	0	66.25	66.33	66.50	66.63	66.00	2
10	50	0	0	74.13	74.25	73.50	73.63	74.00	2
11	22	1	0	65.94	65.94	65.25	65.75	66.00	3
12	22	0	0	66.75	66.69	68.94	69.00	66.50	3
13	22	0	0	73.25	73.13	72.75	73.50	73.00	3
14	23	0	0	72.00	71.75	75.50	75.44	72.00	3
15	22	1	0	64.19	64.00	64.44	64.75	64.50	3
16	54	0	0	71.13	71.06	72.06	72.00	71.63	4
17	53	1	0	63.69	63.63	60.25	60.13	64.00	4
18	50	1	0	66.88	66.94	67.06	67.50	67.50	4
19	52	1	0	66.81	66.75	71.94	71.75	67.00	4
20	48	0	0	72.06	72.19	73.88	74.00	73.00	4
21	23	1	0	66.88	66.88	66.88	66.88	66.00	5
22	21	0	1	69.63	69.88	70.88	70.63	69.00	5
23	23	1	*	65.25	65.50	67.00	66.75	66.00	5
24	22	0	1	73.25	73.50	76.63	76.38	75.00	5
25	23	0	1	67.88	68.13	71.75	71.25	68.00	5
26	51	1	0	63.50	63.75	63.25	63.50	63.50	6
27	52	0	0	75.25	75.50	78.25	78.75	74.00	6
28	36	1	0	65.50	65.75	67.50	67.50	67.00	6
29	45	1	0	64.00	64.00	66.00	66.00	63.50	6
30	22	0	0	67.00	67.25	69.00	69.00	67.00	6
31	22	0	0	69.50	69.25	70.50	71.00	68.00	7
32	22	0	0	72.50	73.00	74.25	74.25	72.00	7
33	63	0	0	67.25	67.50	69.00	68.50	69.00	7
34	59	1	*	60.00	60.25	58.00	58.50	60.00	7
35	23	0	0	76.25	76.50	79.50	79.50	77.00	7
36	58	1	0	61.50	60.94	63.00	63.25	61.00	8
37	36	0	0	70.69	70.75	71.75	70.94	71.00	8
38	31	1	0	63.63	63.81	64.38	64.38	64.00	8
39	31	0	0	69.00	68.94	72.13	71.38	70.00	8
40	28	0	0	67.44	67.75	68.63	68.88	68.00	8
41	48	0	0	71.31	71.50	73.44	73.63	71.00	9
42	65	0	0	67.63	67.56	69.13	69.25	68.00	9
43	26	0	0	75.56	75.38	77.88	77.50	76.00	9

44	60	1	0	64.50	64.56	65.06	65.06	65.50	9
45	52	1	0	64.53	64.63	64.88	64.81	66.00	9
46	44	0	0	69.75	69.88	70.06	70.13	70.00	10
47	24	0	0	69.25	69.50	70.00	70.13	69.50	10
48	43	1	0	63.75	63.63	65.25	65.13	64.00	10
49	35	0	0	78.50	78.25	79.00	79.50	79.00	10
50	46	0	0	69.13	69.19	69.06	69.13	69.00	10
51	40	1	0	59.25	59.56	57.75	57.25	61.00	11
52	30	0	0	69.50	69.25	70.25	70.00	68.00	11
53	26	1	0	61.50	61.56	59.88	60.00	62.00	11
54	57	0	0	71.25	70.50	72.25	71.56	72.00	11
55	28	1	0	62.25	62.75	61.56	61.75	62.00	11
56	45	1	*	65.50	65.50	64.50	65.00	65.00	12
57	24	0	*	71.75	72.00	74.50	74.50	74.00	12
58	22	1	*	64.31	64.31	65.19	65.19	64.00	12
59	54	0	*	68.50	68.50	69.75	69.75	69.50	12
60	21	0	*	67.50	67.50	70.50	70.50	68.00	12
61	44	0	0	70.00	70.00	67.00	67.25	70.00	13
62	42	1	0	63.00	63.00	61.00	61.00	62.00	13
63	24	0	0	69.38	69.38	68.13	68.19	70.00	13
64	27	1	0	62.69	62.75	65.50	65.50	64.75	13
65	26	0	0	64.19	64.25	65.00	65.00	64.00	13
66	48	0	1	70.75	71.00	71.00	70.50	70.00	14
67	20	1	1	64.50	65.00	65.00	64.25	64.00	14
68	42	1	*	61.75	62.00	62.00	61.75	61.00	14
69	24	1	1	67.50	68.00	68.00	68.25	67.00	14
70	20	1	1	63.50	64.00	64.00	63.75	62.00	14
71	25	1	0	64.00	64.00	64.00	64.00	62.00	15
72	39	0	0	72.25	72.25	70.50	70.50	71.75	15
73	31	0	0	66.50	66.50	67.50	67.50	67.00	15
74	58	0	0	70.50	70.50	72.00	72.00	71.50	15
75	58	1	0	58.50	59.00	59.50	59.50	59.00	15
76	25	1	0	73.25	73.25	73.88	73.88	74.00	16
77	48	0	0	72.00	72.13	77.25	77.25	73.00	16
78	51	1	0	67.13	67.38	66.13	66.25	67.00	16
79	25	0	0	71.75	71.75	72.75	72.88	71.00	16
80	44	0	0	72.00	71.88	71.25	71.50	72.00	16
81	46	0	*	69.50	69.50	73.00	73.00	69.00	17
82	25	1	0	66.38	66.50	68.00	68.50	67.00	17
83	47	1	1	65.00	65.13	68.13	68.06	64.00	17
84	29	1	1	63.50	63.50	65.50	65.50	64.00	17
85	26	1	1	63.50	63.50	65.17	65.13	63.00	17
86	23	1	0	64.50	65.00	67.00	66.50	66.00	18
87	28	1	0	70.75	70.75	70.00	69.25	71.00	18
88	29	1	1	65.50	65.25	65.50	65.50	65.00	18
89	48	1	1	63.75	63.25	64.75	64.00	64.00	18

90	25	0	1	66.75	66.50	66.25	66.00	66.00	18
91	21	0	0	71.00	71.25	69.25	69.25	71.00	19
92	24	1	0	64.25	64.50	64.25	64.50	64.00	19
93	54	1	0	66.50	66.50	66.25	66.00	66.00	19
94	44	1	0	66.00	66.00	66.50	67.00	66.00	19
95	46	1	0	63.00	63.00	59.00	58.50	64.00	19
96	45	1	0	65.38	65.75	65.00	64.75	66.00	20
97	61	0	0	67.50	67.50	68.31	69.25	67.50	20
98	50	1	0	60.44	60.75	65.25	65.50	60.25	20
99	22	0	0	70.00	70.50	72.50	73.38	71.00	20
100	23	0	0	68.56	68.81	72.75	73.00	68.50	20
101	26	1	0	64.19	64.13	64.19	63.50	63.00	21
102	32	0	1	69.00	68.50	63.00	63.50	70.00	21
103	24	0	1	78.31	79.13	83.19	84.44	79.00	21
104	22	0	1	74.06	74.19	76.25	76.50	73.00	21
105	22	1	1	71.31	71.25	73.19	73.13	70.00	21
106	23	1	0	64.50	64.50	63.50	63.25	64.00	22
107	52	1	0	68.50	68.00	66.00	66.00	69.00	22
108	49	0	0	71.25	71.50	69.50	69.50	70.00	22
109	21	0	0	68.25	68.00	67.80	67.80	69.00	22
110	23	1	0	61.75	62.00	60.75	60.75	61.50	22
111	45	1	0	65.00	65.00	65.13	65.13	65.00	23
112	52	1	0	60.94	60.94	60.31	60.81	60.00	23
113	23	0	0	68.31	68.50	69.38	69.38	70.00	23
114	26	1	0	63.25	63.25	63.25	63.25	62.00	23
115	45	0	0	68.00	68.00	70.00	70.00	69.50	23
116	47	1	0	67.00	67.00	67.50	67.50	67.00	24
117	22	1	*	65.00	64.50	63.50	63.50	64.00	24
118	22	0	0	70.75	70.75	70.50	70.75	70.00	24
119	48	0	0	73.50	73.38	74.25	74.25	74.00	24
120	21	1	0	65.25	65.19	64.69	64.88	65.50	24
121	41	1	0	65.50	65.44	65.00	65.00	66.00	25
122	51	1	0	64.75	64.75	63.50	63.50	65.00	25
123	26	0	0	73.50	73.50	75.25	75.25	74.00	25
124	24	0	0	71.25	71.25	73.63	73.56	72.00	25
125	56	0	0	67.38	67.38	68.00	68.00	68.00	25
126	40	1	0	65.75	65.75	66.25	66.13	66.00	26
127	42	0	0	70.75	70.75	71.25	71.44	70.00	26
128	51	0	0	73.75	73.75	72.50	72.50	72.00	26
129	42	1	0	64.50	64.56	66.50	66.50	65.00	26
130	45	0	0	69.50	69.75	72.75	72.69	70.00	26
131	22	1	0	59.75	59.75	59.75	60.25	60.00	27
132	54	0	0	66.88	66.50	68.00	68.00	67.00	27
133	52	1	0	61.00	61.50	60.00	60.00	62.00	27
134	57	0	0	71.00	70.50	70.00	70.25	71.00	27
135	51	1	0	60.50	60.25	60.25	61.00	61.00	27

136	33	0	0	77.16	77.25	77.75	77.75	77.00	28
137	32	1	0	58.50	58.62	58.50	58.50	59.00	28
138	30	1	0	59.75	59.45	60.00	59.25	60.00	28
139	21	1	0	66.50	66.25	66.75	67.00	66.00	28
140	29	0	0	71.50	71.50	72.00	72.16	72.00	28
141	25	1	0	61.00	61.00	61.50	61.50	60.25	29
142	62	0	0	67.25	67.25	72.25	72.25	68.00	29
143	29	1	0	65.25	65.50	66.50	66.50	67.00	29
144	53	0	0	69.75	69.75	72.00	72.50	69.00	29
145	50	1	0	64.75	64.75	68.00	68.00	65.00	29
146	22	0	*	69.31	69.25	70.50	71.00	70.00	30
147	23	1	0	64.00	64.25	66.00	66.31	65.00	30
148	22	0	*	67.19	68.25	68.88	68.69	69.00	30
149	23	0	0	70.00	70.19	71.13	71.31	70.00	30
150	24	1	0	62.50	62.00	63.13	63.06	63.00	30
151	22	1	0	65.25	65.25	70.25	70.25	66.00	31
152	23	0	1	68.50	68.50	74.00	74.00	69.00	31
153	37	1	1	63.75	63.75	69.00	69.00	64.00	31
154	22	1	0	64.25	64.25	69.00	68.75	64.00	31
155	23	0	*	65.00	65.25	69.50	69.75	65.00	31
156	32	0	0	71.50	71.25	75.25	75.00	72.00	32
157	48	1	0	66.69	66.69	68.25	68.25	67.00	32
158	44	1	0	63.88	63.88	66.25	66.25	63.00	32
159	60	1	0	62.50	62.63	64.75	64.81	62.00	32
160	40	1	1	65.31	65.38	67.38	67.44	66.00	32
161	21	0	0	67.00	66.75	69.00	69.50	67.00	33
162	21	0	0	66.00	66.50	55.50	55.75	67.00	33
163	22	0	0	71.25	71.50	68.50	68.75	72.00	33
164	21	0	0	70.00	69.00	70.00	70.50	69.00	33
165	21	0	0	70.00	70.25	72.00	71.50	69.00	33
166	24	1	0	67.25	67.13	66.75	67.00	67.00	34
167	23	0	0	69.50	69.50	69.25	69.00	69.00	34
168	20	0	0	71.13	71.13	70.50	70.50	71.00	34
169	51	1	0	65.00	65.00	64.00	63.50	65.00	34
170	53	0	0	71.75	72.00	72.00	72.00	71.00	34
171	22	0	*	69.31	69.25	70.50	71.00	70.00	35
172	23	1	0	64.00	64.25	66.00	66.31	65.00	35
173	22	0	*	67.19	68.25	68.88	68.69	69.00	35
174	23	0	0	70.00	70.19	71.13	71.31	70.00	35
175	24	1	0	62.50	62.00	63.13	63.06	63.00	35
176	22	1	1	65.50	65.50	66.00	66.00	68.00	36
177	35	1	0	69.00	69.00	68.55	68.55	68.50	36
178	47	1	0	68.50	68.50	68.25	68.50	69.00	36
179	24	1	0	69.75	69.50	68.50	68.50	66.00	36
180	34	0	0	70.50	70.00	69.00	69.00	71.00	36
181	27	0	0	69.50	69.50	72.38	72.38	70.00	37

182	27	0	0	70.00	70.00	72.00	72.00	71.00	37
183	49	0	0	68.75	68.75	73.75	73.75	70.00	37
184	48	0	0	72.00	72.00	77.25	77.25	72.00	37
185	51	1	0	66.13	66.13	66.13	66.13	66.00	37
186	57	1	0	65.00	65.25	66.50	66.25	66.00	38
187	25	1	0	66.25	66.50	68.25	68.25	68.00	38
188	59	0	0	66.25	66.50	66.00	66.25	68.00	38
189	24	1	0	68.00	68.00	69.00	69.12	69.00	38
190	22	1	0	69.00	69.25	73.25	73.00	69.00	38
191	24	0	0	67.25	67.00	68.00	67.50	67.00	39
192	23	0	0	68.25	68.38	71.75	70.50	69.50	39
193	22	1	0	64.88	64.75	64.00	64.50	65.00	39
194	36	1	0	63.50	63.25	61.50	62.00	61.00	39
195	25	0	0	71.50	72.00	74.25	74.00	71.00	39
196	65	1	0	60.63	60.63	60.75	60.88	61.00	40
197	57	1	0	61.75	61.75	62.00	62.13	62.00	40
198	50	0	0	72.75	72.75	72.88	72.88	73.00	40
199	22	0	0	75.50	75.50	75.13	75.13	75.00	40
200	23	1	0	65.25	65.38	65.13	65.00	65.00	40
201	50	0	0	66.19	66.19	69.38	68.50	67.00	41
202	50	1	*	69.19	69.31	71.25	71.25	68.00	41
203	26	1	0	70.31	70.75	72.75	72.31	71.00	41
204	23	1	1	62.25	62.19	63.06	63.13	62.00	41
205	40	0	*	70.25	69.50	73.38	73.00	71.00	41
206	24	0	0	70.06	70.06	70.94	71.00	70.00	42
207	26	0	0	69.38	69.38	69.56	69.56	69.50	42
208	24	1	0	64.06	64.06	64.25	64.19	64.00	42
209	50	0	0	72.06	71.94	71.88	72.00	71.50	42
210	50	1	0	60.50	60.44	60.38	60.19	61.00	42
211	22	0	0	71.00	71.00	75.25	75.25	72.00	43
212	23	1	0	65.50	65.50	66.50	67.00	66.00	43
213	22	1	*	63.00	63.00	63.50	63.13	64.00	43
214	23	1	1	62.50	62.06	66.63	67.75	63.50	43
215	22	1	0	60.50	60.50	61.63	61.63	60.00	43
216	36	1	0	63.75	63.00	65.00	65.25	64.00	44
217	41	1	0	65.00	63.50	64.88	66.00	64.00	44
218	35	1	0	66.88	66.50	67.50	67.75	66.00	44
219	41	1	0	67.00	66.00	65.00	63.50	65.00	44
220	26	1	0	62.75	63.00	60.63	61.00	63.00	44
221	30	1	0	66.25	66.50	68.25	68.00	65.00	45
222	34	1	0	64.50	64.75	65.00	65.00	62.50	45
223	49	1	0	61.00	60.50	61.00	61.25	61.00	45
224	48	1	0	69.25	69.00	68.50	68.50	69.00	45
225	30	0	1	69.25	69.50	71.00	71.00	71.00	46
226	42	1	1	63.75	63.75	68.00	68.00	65.50	46
227	29	0	1	68.50	68.75	73.50	73.50	70.00	46

228	31	1	1	60.50	60.75	64.25	64.00	60.50	46
229	64	1	1	61.00	61.00	64.25	64.25	63.00	46
230	29	0	1	70.06	70.06	74.38	74.13	70.00	47
231	33	0	0	71.06	71.75	73.75	74.00	72.00	47
232	43	1	1	65.50	65.38	69.44	69.31	65.50	47
233	48	1	1	62.38	62.50	64.25	64.75	62.50	47
234	62	1	1	63.13	63.25	65.38	65.63	62.75	47
235	25	1	0	67.50	67.88	66.38	66.25	69.00	48
236	22	0	0	71.13	71.00	70.00	70.25	71.00	48
237	40	1	0	65.50	65.75	65.25	65.63	65.00	48
238	43	0	0	73.50	73.25	74.00	74.13	73.00	48
239	37	0	*	68.75	68.75	69.00	69.13	69.25	48
240	23	0	*	69.31	69.25	71.00	71.20	69.00	49
241	24	0	1	66.00	66.00	72.50	72.40	66.00	49
242	23	1	1	62.00	61.87	67.50	67.38	62.00	49
243	36	0	1	72.00	72.50	75.50	75.44	73.00	49
244	29	0	1	68.81	68.75	74.44	75.00	69.50	49
245	39	0	0	63.75	63.75	64.00	64.63	64.00	50
246	51	1	0	60.00	59.50	60.00	59.63	60.00	50
247	23	0	0	70.25	70.13	70.88	72.00	70.00	50
248	50	0	0	69.00	69.13	69.75	69.88	69.00	50
249	22	1	0	64.50	64.50	63.75	63.75	64.50	50
250	22	0	*	68.00	68.25	71.31	71.25	68.00	51
251	21	1	*	63.94	63.81	65.75	65.56	64.00	51
252	30	1	0	58.25	58.25	59.63	59.50	59.00	51
253	48	1	*	63.63	63.25	65.00	65.13	63.00	51
254	54	1	*	65.00	64.88	68.25	68.13	64.00	51
255	60	1	0	64.00	64.00	64.50	64.00	65.50	52
256	38	1	0	68.00	68.00	67.00	67.50	67.50	52
257	23	0	0	71.00	71.00	71.50	71.50	71.00	52
258	31	1	0	62.00	62.00	62.25	62.25	62.00	52
259	22	1	0	70.75	70.75	70.50	70.50	70.00	52
260	25	1	0	63.50	63.50	65.00	65.00	62.00	53
261	53	1	0	64.63	64.50	63.00	63.00	64.00	53
262	25	0	0	71.38	71.63	71.25	71.19	71.00	53
263	33	0	0	69.88	70.50	72.56	72.56	71.00	53
264	22	0	0	69.75	69.88	72.00	71.63	71.00	53
265	26	0	0	70.50	69.50	72.00	73.00	69.00	54
266	54	0	0	75.88	75.88	80.00	80.25	77.00	54
267	51	1	0	66.88	66.63	67.38	67.38	67.00	54
268	20	1	0	65.13	65.13	64.50	64.38	66.00	54
269	62	1	0	65.13	65.13	65.13	65.13	64.00	54
270	41	1	0	61.00	61.06	61.00	61.06	61.00	55
271	37	1	0	68.37	68.43	69.00	69.50	69.00	55
272	34	1	0	69.37	69.25	72.18	72.18	69.00	55
273	52	1	0	61.00	61.00	61.00	61.00	61.00	55

274	37	1	0	68.12	68.06	68.50	68.30	68.50	55
275	25	1	0	65.19	65.25	67.25	67.19	65.00	56
276	23	0	0	69.44	69.44	72.25	72.50	69.50	56
277	45	1	0	62.44	62.38	63.50	63.63	63.00	56
278	25	1	0	61.25	61.25	61.88	60.88	57.00	56
279	47	0	0	71.19	71.19	72.88	72.88	73.00	56
280	22	1	1	61.50	61.50	64.50	64.50	63.00	57
281	28	1	0	67.50	67.50	69.50	69.50	69.00	57
282	20	0	0	64.00	64.00	66.00	66.00	64.00	57
283	26	0	0	73.00	73.00	78.00	78.00	73.00	57
284	23	0	1	67.50	67.50	73.50	73.50	68.00	57
285	25	0	*	68.00	67.94	72.25	72.50	69.00	58
286	21	1	*	65.56	65.50	68.25	68.13	66.00	58
287	22	1	*	65.13	65.31	65.38	65.44	65.00	58
288	33	1	0	64.25	64.38	65.56	65.88	63.00	58
289	22	1	*	65.19	65.44	64.81	65.19	65.00	58
290	42	0	0	68.56	68.69	69.25	69.19	69.00	59
291	44	0	0	72.38	72.31	72.50	72.50	71.50	59
292	40	1	0	61.19	61.25	63.50	63.75	60.00	59
293	40	1	0	67.19	67.00	67.25	67.19	67.50	59
294	40	1	0	64.50	64.56	63.75	63.00	64.25	59
295	25	0	0	67.50	67.50	67.50	67.50	68.00	60
296	31	0	0	71.69	71.69	69.38	69.38	71.00	60
297	52	0	0	66.75	66.75	64.75	64.75	66.00	60
298	50	1	0	66.75	66.75	66.00	66.00	67.00	60
299	44	0	0	65.38	65.38	65.00	65.00	65.00	60
300	23	0	1	72.19	72.06	74.63	74.81	72.00	61
301	22	1	0	63.31	63.19	63.69	63.81	63.50	61
302	44	0	0	73.88	73.81	74.13	74.06	74.00	61
303	43	1	0	62.56	62.63	63.50	63.44	62.00	61
304	47	1	1	63.00	63.13	65.19	65.19	63.00	61
305	23	0	*	71.00	71.00	73.00	72.75	71.00	62
306	52	1	*	62.75	63.00	64.13	64.00	63.00	62
307	21	1	0	61.75	61.75	62.13	62.25	62.50	62
308	30	1	0	58.00	58.25	59.00	59.25	59.00	62
309	49	0	0	70.13	70.25	71.13	71.00	70.00	62
310	45	1	0	56.88	56.88	59.75	59.63	60.00	63
311	21	1	0	61.13	61.13	64.50	64.56	64.00	63
312	24	0	0	68.75	68.75	69.00	69.13	74.00	63
313	56	0	0	68.50	68.50	71.00	70.88	73.00	63
314	22	0	0	65.00	64.88	71.50	71.00	67.00	63
315	32	0	0	73.50	75.30	76.80	76.50	73.00	64
316	30	0	0	69.30	69.50	72.00	73.00	70.00	64
317	29	0	0	66.50	67.00	66.50	67.00	68.00	64
318	48	0	0	68.00	68.00	72.00	72.30	68.00	64
319	48	1	0	70.00	70.00	69.30	69.00	68.00	64

320	24	0	0	72.00	71.00	70.88	70.75	72.00	65
321	44	1	0	62.63	62.50	63.38	63.50	64.00	65
322	42	0	0	74.00	73.00	72.00	72.00	73.50	65
323	47	0	0	71.00	71.13	71.50	71.00	69.00	65
324	22	1	0	64.88	64.88	62.50	62.25	62.75	65
325	28	0	0	72.13	72.00	74.38	74.50	72.00	66
326	48	1	1	66.00	66.00	67.00	67.00	67.00	66
327	47	1	0	62.00	62.00	63.00	63.00	62.00	66
328	24	1	1	64.38	64.38	68.38	68.38	66.00	66
329	22	1	1	65.38	65.38	68.50	68.38	67.00	66
330	26	0	1	69.50	69.50	68.25	68.25	69.50	67
331	20	0	0	66.00	66.00	67.13	67.00	66.00	67
332	48	0	0	70.56	70.31	68.00	68.00	70.00	67
333	22	1	0	66.31	66.44	66.25	66.25	67.00	67
334	60	1	0	64.00	64.00	65.50	65.50	65.00	67
335	22	1	0	60.31	60.44	59.13	58.88	60.00	68
336	56	0	0	65.50	65.69	67.38	66.50	66.50	68
337	50	1	0	65.38	65.38	65.63	66.38	66.00	68
338	22	1	0	64.38	64.75	66.13	66.00	65.00	68
339	22	1	0	65.56	66.25	66.06	66.44	66.00	68
340	24	1	0	67.25	67.13	66.75	67.00	67.00	69
341	23	0	0	69.50	69.50	69.25	69.00	69.00	69
342	20	0	0	71.13	71.13	70.50	70.50	71.00	69
343	51	1	0	65.00	65.00	64.00	63.50	65.00	69
344	53	0	0	71.75	72.00	72.00	72.00	71.00	69
345	24	0	1	66.00	66.00	68.30	68.30	67.00	70
346	35	1	0	67.00	67.00	67.00	67.00	67.00	70
347	22	1	0	65.00	65.00	68.00	68.00	66.00	70
348	23	0	1	68.50	69.00	74.60	74.50	69.00	70
349	40	0	1	66.00	66.00	70.70	70.60	66.00	70
350	21	1	0	66.59	66.59	66.50	66.50	66.50	71
351	23	1	*	61.00	61.00	60.10	60.10	60.50	71
352	51	1	0	67.00	67.00	66.75	66.75	66.25	71
353	24	1	0	64.00	64.00	63.50	63.50	63.00	71
354	53	0	0	70.25	70.25	74.88	74.88	70.00	71
355	22	1	0	62.19	62.06	65.00	65.00	63.50	72
356	25	0	0	68.88	68.88	68.75	69.00	69.00	72
357	47	1	0	62.19	62.25	63.25	63.50	62.00	72
358	51	0	0	68.88	68.88	70.25	70.13	69.00	72
359	50	1	0	67.63	67.63	66.74	66.25	68.00	72
360	45	0	0	68.00	67.75	69.00	69.00	70.00	73
361	25	0	0	70.13	70.13	68.06	69.50	72.00	73
362	21	0	0	73.13	73.75	76.25	76.25	74.00	73
363	46	1	0	63.50	63.50	61.00	61.50	64.00	73
364	40	1	0	66.38	66.38	64.50	64.50	66.00	73
365	37	1	1	63.50	63.50	68.80	68.80	63.50	74

366	28	0	1	66.50	66.00	66.80	66.80	66.50	74
367	29	1	1	61.50	61.50	61.50	61.50	63.50	74
368	27	1	1	63.50	63.70	67.50	67.50	62.50	74
369	28	0	1	66.50	66.80	66.80	66.50	67.50	74
370	24	1	0	63.00	63.00	66.00	66.00	64.00	75
371	23	1	0	67.50	68.00	68.50	68.50	68.00	75
372	26	0	0	71.00	71.00	74.00	74.00	73.00	75
373	25	0	*	70.00	70.00	73.00	73.50	71.00	75
374	23	1	1	69.00	69.50	74.00	74.50	68.00	75
375	28	0	1	68.38	68.38	72.00	72.13	70.00	76
376	36	0	0	68.88	68.75	70.50	71.25	70.00	76
377	31	0	0	71.25	71.31	76.00	76.13	72.00	76
378	53	1	0	62.63	62.75	64.38	63.75	63.00	76
379	27	1	*	64.38	64.25	68.25	67.75	65.00	76
380	23	1	1	63.00	62.25	64.50	64.75	63.00	77
381	37	0	1	68.56	68.63	73.00	73.00	70.50	77
382	23	0	1	64.50	64.50	67.00	67.00	64.00	77
383	23	0	1	70.75	70.75	71.06	71.06	71.00	77
384	26	1	1	64.50	64.50	64.00	64.00	64.50	77
385	22	0	*	69.75	69.87	69.87	70.00	70.00	78
386	21	1	0	66.87	66.75	66.75	66.75	67.00	78
387	22	0	0	68.12	68.37	68.25	68.12	68.00	78
388	22	1	1	64.12	64.12	64.25	64.12	64.00	78
389	22	1	0	65.12	65.12	65.12	65.00	65.00	78
390	61	1	0	60.13	61.13	61.13	61.38	61.00	79
391	39	1	0	62.50	62.50	65.25	65.63	63.00	79
392	55	1	0	60.63	60.56	64.00	64.13	62.00	79
393	55	1	1	59.25	59.50	63.75	63.88	59.50	79
394	38	1	0	68.50	68.88	69.25	69.00	68.50	79
395	58	1	0	65.00	65.25	66.75	67.00	66.50	80
396	50	0	0	73.50	73.00	75.75	76.00	73.00	80
397	19	1	0	66.75	67.00	68.75	69.00	67.00	80
398	60	1	*	60.00	60.00	61.00	61.00	61.00	80
399	67	1	*	61.00	61.25	64.00	64.50	62.00	80
400	46	1	0	58.13	58.13	57.50	57.25	58.50	81
401	51	1	0	66.50	66.50	70.00	69.88	67.00	81
402	52	0	0	71.50	71.50	72.50	72.50	72.50	81
403	47	0	0	73.13	73.00	74.50	74.50	73.00	81
404	21	0	0	68.00	68.00	68.50	68.25	69.00	81
405	22	1	0	62.50	62.75	63.50	63.00	63.00	82
406	22	1	0	66.00	66.50	62.50	62.75	65.00	82
407	22	0	0	70.75	70.00	71.50	71.00	71.50	82
408	23	1	0	65.75	65.50	64.75	64.75	65.00	82
409	47	1	0	61.00	61.50	62.50	63.75	61.00	82

Legend: id = subject identification number
age = age in years at time of study
sex = 0 if male, 1 if female
race = 0 if white, 1 if black
ht1 = first measured height (in inches)
ht2 = second measured height (")
arm1 = first measured arm span (")
arm2 = second measured arm span (")
self = self-reported height (")
meas = measurer identification number
* = missing data (for subjects of races other than white and black)

Note: The height and arm span measurements are reported to the nearest hundredth of an inch. For example, Subject 1's height measurements were 61 7/8 inches (decimal equivalent = 61.875, rounded to 61.88) and 61 1/2 inches (decimal equivalent = 61.50); and her arm span measurements were 62 1/4 inches (decimal equivalent = 62.25) and 62 3/8 inches (decimal equivalent = 62.375, rounded to 62.38). Some of the measurements have been incorrectly reported and/or rounded (see text).

APPENDIX B: The validity of measuring instruments

The validity of a measuring instrument is usually defined as the extent to which an instrument measures "what you want it to measure" or "what it purports to measure". (I've never heard the word "purport" used in any other context. Have you?) Validity is concerned with relevance, just as reliability is concerned with consistency.

In the old days (when I was starting out in the measurement and statistics business in the early sixties, and prior to that time as well) people talked about three kinds of validity: (1) content validity--the extent to which a measuring instrument "covered" the domain for which it was intended; (2) criterion-related validity--the extent to which the measurements obtained with a given instrument agreed (relatively, but not necessarily absolutely) with some external "gold-standard" criterion; and (3) construct validity--the extent to which the measurements obtained with a given instrument agreed with theoretical expectations. The term "face validity" was sometimes used as synonymous with content validity, but usually referred to the extent to which the "objects" (people) that were measured with the instrument agreed that the instrument was valid. Criterion-related validity was further broken down into concurrent validity and predictive validity, depending upon whether the gold-standard measurements were taken at approximately the same time as the measurements for the instrument whose validity was in question, or were taken at a later point in time (thus the "predictive" label). Construct validity was similarly broken down into convergent validity and discriminant (sometimes called divergent) validity--see, for example, Campbell and Fiske (1959).

Shortly thereafter, discussions of several additional kinds of validity appeared in the literature, largely due to the unfortunate (in my opinion) choice of the terms "internal validity" and "external validity" by Campbell and Stanley (1966) and by Cook and Campbell (1979) in their now-classic treatises on experimental design (those two terms have little or nothing to do with validity in the measurement sense of the term).

More recently, the term "construct validity" has been alleged to subsume all of the other kinds of measurement validity (see Messick, 1989 and elsewhere), and it has taken a pre-eminent (although controversial) position in the literature. The term has become almost synonymous with the term "science" (in my opinion) and has consequently lost (again, in my opinion) its formerly admirable specificity. Advocates of the pre-eminence of construct validity even proclaim that an investigation of an instrument's validity must include evidence regarding the consequences of its use. (That proclamation, frankly, blows my mind. Could a yardstick that is alleged to measure height really be declared invalid if a person who is measured with it does not make her(his) school's basketball team

because (s)he is declared to be too short, and as a result loses interest in all physical exercise?)

My personal position is that validity ultimately, if not immediately, boils down to a matter of expert judgment--content validity, if you will. (Who qualifies as an expert may be difficult to determine, but see Weiss & Shanteau, 2003a and 2003b for one possible approach). Those who advocate correlating measurements obtained for a given instrument with those obtained with a gold-standard instrument must either assume that the gold-standard instrument has itself been declared to be valid by expert judgment or be willing to acknowledge a situation of infinite regress--correlating the gold-standard measures with platinum-standard measures, for example?! (See Ebel, 1961 regarding this same “infinite regress” argument, and see Wacholder, Armstrong, & Hartge, 1993 for an interesting discussion of the use of an "alloyed" gold standard in epidemiological research.) Comparing obtained measurements with theoretical expectations is commendable, but if the two disagree how do we know whether it is the instrument or the theory, or both, that is at fault? (Things aren't much better if the two agree. Again, both could be wrong.)

In Chapter 4 of this book I discussed the concept of attenuation. One of the most common applications of the correction for attenuation is to the correlation between scores obtained for a measuring instrument whose validity is being investigated and scores obtained for an external criterion that serves as a gold standard. The simplest case is the correlation between aptitude test scores and subsequent achievement scores. If an aptitude test is to be valid that correlation must be high and positive. (Those who obtain higher aptitude test scores must also obtain higher achievement test scores, by definition of the concept of aptitude.) Suppose a validity study of the Scarlet Aptitude Test (I'm making up these names) yielded the data that I cited earlier in Chapter 4, i.e., an obtained correlation of .54, a reliability coefficient of .64 for the aptitude test, and a reliability coefficient of .81 for the Gold Achievement Test. Application of the correction for attenuation formula would produce an estimate for the true correlation of .75. So far, so good. Now suppose that a validity study of a competing instrument, the Gray Aptitude Test, yielded the same obtained correlation with the scores obtained on the Gold Achievement Test, .54, a reliability coefficient of .49 for that aptitude test, and the same reliability coefficient, .81, for the achievement test. The correction for attenuation formula would produce an estimate for the true correlation of .86. This would seem to indicate that you could get better validity (stronger estimated correlation between true aptitude and true achievement) by using the less reliable test!

There is also something called “the attenuation paradox” (see, for example, Loevinger, 1954), where it is alleged that if you increase internal consistency reliability to the point where all of the items on a test correlate perfectly with one another so that there are only two possible total scores, 0 and k (k is the number of items on the test) the validity of the test may actually decrease, contrary to the

expectation provided by the “correction for attenuation” formula. Humphreys (1956) claimed, however, that the paradox vanishes when you give up the normality assumption for the gold-standard criterion variable. [See also Feldt, 1997 for a discussion of reliability and validity going in opposite directions.]

Many of the same people who object to talking about the reliability of an instrument (rather than the reliability of scores obtained by using the instrument) also object to talking about the validity of an instrument. They insist that we should refer to the validity of the measurements, i.e., the obtained scores, or to the validity of interpretations made concerning the measurements. I respectfully disagree and believe that this is making much ado about nothing. Of course validity will vary from time to time, from study to study, etc., just like reliability does, but if you are concerned with validity for this object (or these objects) on this occasion (or these occasions), no great harm is done by attaching the term "validity" to the instrument itself. For more on the "evolving" concept of validity, see the article by Borsboom, Mellenbergh, and Van Heerden (2002) and Chapter 6 of Boorsboom's text (2005).

In the next three appendices I include a discussion of the reliability and the validity of birth certificates and death certificates, measurements of height and weight, and the gospels of Matthew, Mark, Luke, and John.

APPENDIX C: The reliability and validity of birth and death certificates

Two examples of measuring instruments that have items but do not have total scores are birth certificates and death certificates. My colleague, Dr. Sally Northam (University of Texas at Tyler), and I have recently written papers on the reliability and validity of those instruments. In this appendix I would like to summarize some of our findings.

The first thing we found is that the terms “reliability” and “validity” are not universally used to indicate the consistency and relevance (respectively) of birth and death certificates. Just as I indicated in Chapter 1 of this book, synonyms for reliability such as “agreement” and synonyms for validity such as “accuracy” are commonly encountered in reports of the reliability and/or validity of the various items on the certificates.

Another finding was that the type of reliability most often studied was inter-rater reliability--the extent to which two or more equal status persons agree with respect to the data recorded on the certificates. The situation regarding validity was less clear, and actually rather confusing. The researchers seemed to be reluctant to proclaim the certificates themselves as the “gold standards” or to designate any other sources (e.g., medical records) as the criteria against which those certificates need to be validated. Curious.

The agreement between raters (or sources of ratings) was generally good, but ranged from less than 50% for month prenatal care began (birth certificate) to approximately 90% for maternal age (also birth certificate). People who fill out birth certificates or death certificates are often poorly trained in the proper completion of such forms. [We discovered that nurses are permitted to certify deaths in certain states. Sally is a nurse, and I was a faculty member of schools or colleges of nursing for 20 years, but we were both surprised to find out that some nurses actually had such responsibilities.]

When an external source was taken as a gold standard, the criterion-related validity ranged from 0 to almost perfect. (The authors of one study of 68 death certificates reported that not a single one of them indicated the correct cause of death; whereas the authors of another study reported a sensitivity of 94.8% for the identification of congenital abnormalities on birth certificates.) Although an autopsy is arguably the most defensible gold standard for determining cause of death, only about 10-15% of deaths are subject to autopsies.

The most serious weakness of death certificates was their under-estimation of numbers of particular kinds of deaths and their under-estimation of particular causes of deaths. Most prominent in the former category was fetal deaths; they do not include spontaneous abortions (miscarriages), legal therapeutic abortions (a very controversial matter), or illegal therapeutic abortions (obviously). Most prominent in the latter category were suicides (also a very controversial matter).

APPENDIX D: The reliability and validity of height and weight measurements

[This section consists of excerpts from my book on height and weight (Knapp, 2004).]

Reliability of height measurements

For a sample of 229 subjects (95 male, 134 female) in NHANES II, Marks, Habicht, and Mueller (1989) found the measure/re-measure correlation (reliability coefficient) of a height stadiometer (that instrument in doctors' offices that you stand on) to be very high (a correlation of approximately .97).

Voss, Bailey, et al. (1990) measured ten children three times each with five different stadiometers. The children's "true" heights ranged from 106.0 cm to 152.0 cm. The average difference from true height was about .2 to .3 cm.

In a study carried out by Rodacki, et al. (2001), ten subjects (five males and five females) had their standing heights measured with a stadiometer 150 times each [wow!], in 3 series of 10 sets of 5 measurements, with "breaks" between the series for the subjects to get off and then back on the stadiometer. (See their article for more details and for some great pictures.) The average discrepancy between one measure and another ("un"reliability) was of an order of magnitude of approximately one-half millimeter. (Again see their article for more details--the analysis was rather complicated--and for a comparable discussion of their findings regarding the measurement of the sitting heights of ten other subjects.)

Dr. Janet Engstrom and her colleagues have carried out several investigations of the reliability of infantometers (and of tape measures--see below) for measuring the supine length of newborns. They concentrate on absolute measures of reliability, such as average discrepancies between corresponding measurements, rather than correlations between two sets of measurements, because, as she (Engstrom, 1988) and others (e.g., Rogosa, 2002; Baker & Kramer, 2003) have argued, you can get a perfect correlation between two sets of measurements, e.g., 1, 2, 3, 4, 5 and 10, 20, 30, 40, 50, yet have very poor agreement between the actual magnitudes of the measurements. Here are some of their findings:

Johnson, et al. (1998): Using the Neo-infantometer for a sample of 32 babies, the within-examiner mean absolute discrepancy (intra-examiner reliability) was .50 cm for one examiner and .71 cm for a second examiner. The between-examiner mean absolute discrepancies (inter-examiner reliability) were .81 cm for the first comparison between the two examiners (Examiner A's first set of measurements vs. Examiner B's first set of measurements) and .61 cm for the second comparison (of their second sets of measurements).

Johnson, et al. (1999): Using the Auto-length™ measuring device for a sample of 48 healthy term infants, the intra-examiner reliabilities for two examiners were .60 cm and .84 cm, respectively; and their inter-examiner reliabilities were 1.02 cm and .82 cm.

For references to some older studies of the reliability of infant stadiometers, see Table 1 in the Johnson, Engstrom, and Delhar (1997) article. And if you're interested in the history of the measurement of infant length, see the excellent article by Johnson and Engstrom (2002).

As I indicated above, some of Engstrom's studies were also concerned with the reliability of tape measures for measuring infant length. For a sample of 48 infants measured twice by each of two registered nurses, Rosenberg, et al. (1992) found a mean absolute difference between first and second measurements of .64 cm for Nurse 1 (intra-measurer reliability), a mean absolute difference of .50 cm for Nurse 2 (also intra-measurer reliability), a mean absolute difference of .89 cm between their first measurements (inter-measurer reliability), and a mean absolute difference of .84 cm between their second measurements (also inter-measurer reliability). Johnson, et al. (1997) reported mean intra-examiner absolute differences of .92 cm and 1.18 cm for two examiners who measured a sample of 50 newborns twice each; the corresponding inter-examiner statistics were .74 cm and .84 cm. Johnson, et al. (1998) reported average intra-examiner discrepancies of .80 cm and .53 cm, and average inter-examiner discrepancies of .74 cm and .84 cm for a sample of 32 babies. And Johnson, et al. (1999) reported intra-examiner reliabilities of .92 cm and .74 cm, and inter-examiner reliabilities of 1.13 cm and 1.39 cm (n = 48).

A first study carried out by Dr. Jean Brown and her colleagues (Brown, Whittemore, & Knapp, 2000) provided some evidence for the reliability of tape measures for measuring height. They found a measure/re-measure reliability coefficient of .998 and a mean absolute difference of .20 cm for the Stanley model 33-158 tape measure. In their second study (Brown, Feng, & Knapp, 2002) the measure/re-measure reliability coefficient was .997 and the mean absolute difference was .43 cm.

Validity of height measurements

Research on the validity of height measurements is rather scarce. A stadiometer looks like it measures height (so-called "face validity") and the experts (the anthropometrists) tell us that it does (a type of content validity). Most investigators who use stadiometers tacitly assume that they are the "gold standards" for measuring height and they see no reason to "validate" them.

One of the best studies of the validity (they call it reliability) of devices for measuring infant length is the research reported by Byrne and Lenz (2002). They compared three instruments: an ordinary cloth tape measure, a portable

device (the Measure Mat), and a traditional stationary infantometer (the Fairgate Rule stadiometer, which seemed to be the gold standard). For the tape measure they found a mean absolute difference (between the tape measure and the infant stadiometer) of 1.75 inches for one sample of 30 infants and a mean absolute difference of 1.05 inches for a second sample of 15 infants. For the Measure Mat the mean absolute differences (between the MeasureMat and the infant stadiometer) were smaller: .92 inches for one sample of 15 infants and .57 inches for a second sample of 15 infants.

In an interesting study of 28 parents' ability to make accurate measurements of their infants' recumbent length (RL) and other anthropometric variables, Bradley, Brown, and Himes (2001) found that for RL the correlation with measurements made by a trained observer was a disappointing .81. (They also called their study a reliability study, but since the measurements provided by a trained, "higher status" observer served as the criteria, it is better designated as a validity study.)

There has been recent research and development work on three-dimensional surface anthropometry that would provide height, weight, and other body measurements "in one fell swoop", so to speak. (See, for example, Jones & Rioux, 1997 and/or any of the other articles in that special issue of Optics and Laser Engineering.) That would make such a device "the new gold standard" for the measurement of height and weight (and body mass index and body surface area, as well as other dimensions that are presently determined one variable at a time.)

Reliability of weight measurements

Empirical investigations of the reliability and the validity of weight scales (top-quality scales found in doctors' offices and ordinary bathroom scales) are even more scarce than their height counterparts. For the "gold standard" scales there appears to be the same acquiescence to the experts and to the manufacturers that such scales are both reliable and valid. (Researchers consider ordinary bathroom scales as inferior for measuring weight as yardsticks are for measuring height, so the former have also not been seriously studied, as far as I can tell.).

The best empirical studies of the reliability of weight scales have been carried out by Engstrom and her colleagues, often in conjunction with their studies of the reliability of instruments for measuring infant length. Johnson, Engstrom, & Delhar (1997) used an electronic scale (Smart Scale Model 20, Olympic Medical; Seattle, WA, U.S.A) to measure the weights of a sample of 50 infants. To quote from their article: "This scale integrates the activity level of an infant by automatically taking 10 weights in rapid succession. A mean of the 10 weights is calculated and displayed as a digital readout. " (p. 500) They found very small mean absolute differences (intra-examiner 1.88 and 3.28 g; inter-examiner 1.94 and 1.66 g).

In an earlier set of three related articles, Kavanaugh, Meier, & Engstrom (1989), Kavanaugh, Engstrom, et al. (1990), and Meier, Lysakowski, et al. (1990) reported the results of a study of the weighing of a sample of 50 infants with two types of scales (a traditional mechanical scale and a an electronic SMART™ scale). They investigated both intra-measurer and inter-measurer reliability. For the former they found a mean absolute difference of 5.50 g for the mechanical scale and a mean absolute difference of 1.36 g for the electronic scale. For the latter they found a mean absolute difference of 18.00 g for the mechanical scale and a mean absolute difference of .88 g for the electronic scale. (As in several of their studies of infant length they also reported the technical error of measurement and other indicators of the reliability of the two types of scales.)

Validity of weight measurements

In two later studies of the measurement of weight for atypical infants, Meier, Engstrom, et al. (1994) and Engstrom, Kavanaugh, et al. (1995) investigated the measurement properties of the BabyWeigh™ electronic scale and two SMART scales (Model 20 and Model 35). The former study was concerned with the validity of the BabyWeigh scale for in-home weighing of 30 preterm and/or high risk infants, using the SMART Model 20 scale as the "gold standard". The mean absolute difference between corresponding measurements for the two scales was 1.30 g. The latter study was concerned with the reliability of the SMART Model 35 scale for the in-bed weighing of 32 critically ill infants. They provided several summary statistics for both intra-measurer and inter-measurer reliability (there were four examiners who took the measurements). The mean absolute intra-measurer difference was 12.58 g for weights obtained in the incubator and was 19.19 g for weights obtained under the radiant warmer; the corresponding mean absolute intermeasurer differences were 14.29 g for incubator and 24.42 g for radiant warmer.

APPENDIX E: The reliability and validity of the four gospels:
A statistical approach

Introduction

There is a vast literature regarding the extent to which the gospels of Matthew, Mark, Luke, and John are historically trustworthy. (See, for example, Cable, no date; Ehrman, 2005; Funk, et al., 1993, 1998; Habermas, 2005; and Roberts, 2007 for a variety of viewpoints.) Most authors use the term "reliable" as synonymous with "true"; others use the term "valid" as synonymous with "true". None, as far as I can determine, explain the important technical difference between reliability and validity; and most offer various substantive, non-statistical arguments as evidence pro or con the truth of those gospels. In what follows I provide some statistical evidence regarding their reliability and their validity. And in so doing I appeal to the analogue of standardized educational testing. (There are a few other statistical analyses of the gospels. See, for example, the linguistic approach by Dave Gentile. His unpublished results can be found at www.davegentile.com/synoptics/main.html.)

Reliability vs. validity

Something is reliable if it is consistent from source to source. Something is valid if it is accurate. My favorite example to illustrate the distinction is the problem: "What is the probability of getting two heads in four tosses of a fair coin?" Most people say 2/4 or 1/2 or .5 or 50%. They are consistent with one another. However, the correct answer is 6/16 or 3/8 or .375 or 37.5%. (Do the math.) Those sources (the people who say 2/4) are reliable but they're not valid.

Standardized tests

Most readers of this paper are familiar with standardized tests such as the Scholastic Aptitude Test (SAT), the Graduate Record Examination (GRE), the Law School Admission Test (LSAT), and the like. Each of those tests has multiple "forms" that have been developed to be "parallel"; i.e., they are interchangeable, so that it doesn't matter which form is given to which examinee at which sitting. In order to determine how parallel they are, they are subjected to several statistical evaluations. Do they yield equal, or approximately equal, average scores? Do scores on the various forms correlate highly? Etc. Those questions are concerned primarily with the reliability of the particular test that is being evaluated. The closer the average scores are on, say, Form A and Form B, and the higher the scores on the two forms correlate with one another, the more reliable they are.

But reliability is not enough. The forms also need to be shown to be valid, i.e., that they are measuring what they are supposed to be measuring. That requires

some external evidence in the form of expert judgment and/or high correlations with already-established criteria. For the SAT, GRE, and LSAT, the evidence is typically provided by subject-matter panels and by the relationship between scores on those tests and the grades obtained in courses that the examinees later pursue.

Similarities between four forms of the SAT and the four gospels

Parallelism of the gospels of Matthew, Mark, Luke, and John has been the focus of much biblical research. There are many books and many easily-accessible internet sources (some free, some requiring a modest fee) that provide chapter-and-verse comparisons of the four gospels. The one I prefer is the one called www.gospelparallels.com, in which the interested person can determine whether or not, or to what extent, a description of a particular event in a passage of one of the gospels is replicated in a passage in one or more of the other gospels. To take a rather well-known example, the Parable of the Sower is found in Matthew (13:1-9), Mark (4:1-9), and Luke (8:4-8), but not in John.

Investigating the reliability of the four gospels is very much like investigating the reliability of four of the forms (say A, B, C, and D) of the SAT. Do the four gospels have the same numbers of "test items" (events to which they refer)? No. (See below.) How well do the "items" in the four gospels agree with one another? Fairly well, except for the well-known discrepancies between John and the other three. (Also see below.) The principal difference between an evaluation of the reliability of the SAT and an evaluation of the reliability of the four gospels is that there are no "scores" for the gospels.

Ah, but how about validity? The validity of the SAT has been controversial ever since the test was first introduced over 80 years ago. Who are the "experts" who say that the test measures scholastic aptitude? Are college grades the appropriate criteria? The validity of the four gospels is equally controversial. Whose external-to-the-New Testament writings can we turn to for evidence? Josephus? Eusebius? Those are tough questions. As you will see, I take a stab at answering such questions, but I will be much more successful in providing evidence for the reliability of the gospels than for their validity.

The gospel parallels and their reliability

The following table, which has been reproduced with the permission of Geoff, webmaster of the www.gospelparallels.com website, tells much of the story regarding the harmony (parallelism) of the four gospels. [I have modified it by eliminating repeated events.]

Table 1: Gospel Harmony

	Matthew	Mark	Luke	John
Prologues	1:1	1:1	1:1-4	1:1-18
The Promise of the Birth of John the Baptist			1:5-25	
The Annunciation			1:26-38	
Mary's Visit to Elizabeth			1:39-56	
The Birth of John the Baptist			1:57-80	
The Genealogy of Jesus	1:2-17		3:23-38	
The Birth of Jesus	1:18-25		2:1-7	
The Adoration of the Infant Jesus	2:1-12		2:8-20	
The Circumcision and Presentation in the Temple			2:21-38	
The Flight into Egypt and Return	2:13-21			
The Childhood of Jesus at Nazareth	2:22-23		2:39-40	
The Boy Jesus in the Temple			2:41-52	
John the Baptist	3:1-6	1:2-6	3:1-6	1:19-23
John's Preaching of Repentance	3:7-10		3:7-9	
John Preaching and Replies to Questioners			3:10-14	
John's Messianic Preaching	3:11-12	1:7-8	3:15-18	1:24-28
The Baptism of Jesus	3:13-17	1:9-11	3:21-22	1:29-34
The Temptation	4:1-11	1:12-13	4:1-13	
The Call of the First Disciples				1:35-51
The Marriage at Cana				2:1-11
The Sojourn at Capernaum				2:12
The First Journey to Jerusalem				2:13
Jesus in Jerusalem (Cleansing the Temple), Return to Bethany	21:10-17	11:15-17	19:45-46	2:14-22
Jesus' Ministry in Jerusalem				2:23-25
The Discourse with Nicodemus				3:1-21
Jesus' Ministry in Judea				3:22
John's Testimony to Christ				3:23-36
The Journey into Galilee	4:12	1:14a	4:14a	4:1-3
The Discourse with the Woman of Samaria				4:4-42
Ministry in Galilee	4:13-17	1:14b-15	4:14b-15	4:43-46a
Jesus' Preaching at Nazareth	13:53-58	6:1-6	4:16-30	
The Call of the Disciples	4:18-22	1:16-20		
Teaching in the Synagogue at Capernaum		1:21-22	4:31-32	

The Healing of the Demoniac in the Synagogue		1:23-28	4:33-37	
The Healing of Peter's Mother-in-law	8:14-15	1:29-31	4:38-39	
The Sick Healed at Evening	8:16-17	1:32-34	4:40-41	
Jesus Departs from Capernaurn		1:35-38	4:42-43	
First Preaching Tour in Galilee	4:23	1:39	4:44	
The Miraculous Catch of Fish			5:1-11	
The Cleansing of the Leper	8:1-4	1:40-45	5:12-16	
The Healing of the Paralytic	9:1-8	2:1-12	5:17-26	
The Call of Levi (Matthew)	9:9-13	2:13-17	5:27-32	
The Question about Fasting	9:14-17	2:18-22	5:33-39	
Plucking Grain on the Sabbath	12:1-8	2:23-28	6:1-5	
The Man with the Withered Hand	12:9-14	3:1-6	6:6-11	
Jesus Heals Multitudes by the Sea	4:24-25 12:15-21	3:7-12	6:17-19	
The Commisioning of the Twelve Apostles	10:1-16	3:13-19 6:7-13	6:12-16 9:1-6	
The Beatitudes	5:1-12		6:20-23	
The Salt of the Earth	5:13	9:49-50	14:34-35	
The Light of the World	5:14-16	4:21	8:16	
On the Law and the Prophets	5:17-20		16:16-17	
On Murder and Anger and Reconciliation	5:21-26		12:57-59	
On Adultery and Divorce	5:27-32		16:18	
On Swearing and Oaths	5:33-37			
On Retaliation	5:38-42		6:29-30	
On Love of One's Enemies	5:43-48		6:27-28 6:32-36	
On Almsgiving	6:1-4			
On Prayer	6:5-6			
The Lord's Prayer	6:7-15		11:1-4	
On Fasting	6:16-18			
On Treasures	6:19-21		12:33-34	
The Sound Eye	6:22-23		11:34-36	
On Serving Two Masters	6:24		16:13	
On Anxiety	6:25-34		12:22-32	
On Judging	7:1-5		6:37-42	
On Profaning the Holy	7:6			
God's Answering of Prayer	7:7-11		11:9-13	
The Golden Rule	7:12		6:31	
The Straight and Narrow and Wide Broad Gate	7:13-14		13:23-24	

The Test of a Good Person, “By their Fruits”	7:15-20		6:43-45	
Not Every One who “Says Lord, Lord” shall Enter into the Kingdom	7:21-23		6:46 13:25-27	
The House Built upon the Rock	7:24-27		6:47-49	
The End and the Effect of the Sermon	7:28-29			
The Woes			6:24-26	
The Centurion of Capernaum and his Servant	8:5-13		7:1-10 13:28-29	4:46b-54
The Widow's Son at Nain			7:11-17	
On Following Jesus, the Would-be Followers	8:18-22		9:57-62	
Stilling the Storm	8:23-27	4:35-41	8:22-25	
The Gadarene (Gerasene) Demoniacs	8:28-34	5:1-20	8:26-39	
Jairus' Daughter and the Woman with a Hemorrhage	9:18-26	5:21-43	8:40-56	
Two Blind Men Healed	9:27-31 20:29-34	10:46-52	18:35-43	
The Dumb Demoniac Healed	9:32-34 12:22-24		11:14-15	
The Harvest is Great	9:35-38		10:2	
The Coming Fate and Persecution of the Disciples	10:17-25 24:9-14	13:9-13	12:11-12 21:12-19	
Exhortation to Fearless Confession	10:26-33		12:2-9	
Divisions within Households	10:34-36		12:51-53	
Conditions of Discipleship	10:37-39		14:25-27	
Rewards of Discipleship	10:40-42		10:16	
Continuation of Journey	11:1			
John the Baptist's Question and Jesus' Answer	11:2-6		7:18-23	
Jesus' Witness concerning John	11:7-19		7:24-35 16:16	
Woes Pronounced on Galilean Cities	11:20-24		10:12-15	
Jesus' Thanksgiving to the Father	11:25-27		10:21-22	
“Come unto Me”	11:28-30			
The Woman with the Ointment	26:6-13	14:3-9	7:36-50	12:1-8
The Ministering Women			8:1-3	
Jesus is Thought to be Beside Himself		3:20-21		
The Sin against the Holy Spirit	12:31-37	3:28-30	12:10	

Against Seeking Signs, the Sign of Jonah	12:38-42 16:1-4	8:11-12	11:16 11:29-32	
The Return of the Unclean Evil Spirit	12:43-45		11:24-26	
Jesus' True Kindred Relatives	12:46-50	3:31-35	8:19-21	
The Parable of the Sower	13:1-9	4:1-9	8:4-8	
The Reason for Speaking in Parables	13:10-17	4:10-12	8:9-10 8:18b 10:23-24	
Interpretation of the Parable of the Sower	13:18-23	4:13-20 4:22-25	8:11-15	
The Parable of the Seed Growing Secretly		4:26-29		
The Parable of the Tares (Weeds)	13:24-30			
The Parable of the Mustard Seed	13:31-32	4:30-32	13:18-19	
The Parable of the Leaven (Yeast)	13:33		13:20-21	
Jesus' Use of Parables	13:34-35	4:33-34		
Interpretation of the Parable of the Tares	13:36-43			
The Parables of the Hidden Treasure and of the Pearl	13:44-46			
The Parable of the Net	13:47-50			
Treasures New and Old	13:51-52			
Second Journey (to Jerusalem)				5:1
The Healing at the Pool called Bethesda				5:2-47
Herod Thinks Jesus is John, Raised	14:1-2	6:14-16	9:7-9	
The Imprisonment and Death of John the Baptist	14:3-12	6:17-29	3:19-20	
The Return of the Apostles		6:30-31	9:10a	
Feeding the Five Thousand	14:13-21	6:32-44	9:10b-17	6:1-15
The Walking on the Water	14:22-33	6:45-52		6:16-21
Healings at Gennesaret	14:34-36	6:53-56		6:22-25
The Bread of Life				6:26-59
What Defiles a Person - Traditional and Real	15:1-20	7:1-23	11:37-41 6:39	
The Syrophoenician (Canaanite) Woman	15:21-28	7:24-30		
Jesus Heals a Deaf Mute and Many Others	15:29-31	7:31-37		
Feeding of the Four Thousand	15:32-39	8:1-10		
The Leaven (Yeast) of the Pharisees	16:5-12	8:14-21	12:1	
A Blind Man is Healed at Bethsaida		8:22-26		

Many Disciples Take Offense at Jesus				6:60-66
Peter's Confession at Caesarea Philippi	16:13-20	8:27-30	9:18-21	6:67-71
Jesus Foretells His Passion	16:21-23	8:31-33	9:22	
“If Any Man would Come after Me”	16:24-28	8:34-9:1	9:23-27	
The Transfiguration	17:1-9	9:2-10	9:28-36	
The Coming of Elijah	17:10-13	9:11-13		
Jesus Heals a Boy Possessed by a Spirit	17:14-18	9:14-29	9:37-43a	
On Faith	17:19-21	9:28-29	17:5-6	
Jesus Foretells His Passion again	17:22-23	9:30-32	9:43b-45	
Payment of the Temple Tax	17:24-27			
True Greatness	18:1-5	9:33-37	9:46-48	
The Strange Exorcist		9:38-41	9:49-50	
Warnings concerning Offenses	18:6-9	9:42-50	17:1-2	
The Parable of the Lost Sheep	18:10-14		15:3-7	
On Reproving One's Brother	18:15-18		17:3	
“Where Two or Three are Gathered Together”	18:19-20			
On Reconciliation	18:21-22		17:4	
The Parable of the Unforgiving Servant	18:23-35			
Decision to Go to Jerusalem	19:1-2	10:1	9:51	
Jesus is Rejected by Samaritans			9:52-56	
The Return of the Seventy			10:17-20	
The Lawyer's Question	22:34-40	12:28-34	10:25-28	
The Parable of the Good Samaritan			10:29-37	
Mary and Martha			10:38-42	
The Importunate Friend at Midnight			11:5-8	
True Blessedness			11:27-28	
Warning against Greed for Wealth			12:13-15	
The Parable of the Rich Fool			12:16-21	
Watchfulness and Faithfulness	24:42-51		12:35-48	
Repentance or Destruction (the Parable of the Barren Fig Tree)			13:1-9	
The Healing of the Crippled Woman on the Sabbath			13:10-17	
A Warning against Herod			13:31-33	
The Lament over Jerusalem	23:37-39		13:34-35	
The Healing of the Man with Dropsy			14:1-6	
Teaching on Humility			14:7-14	

The Parable of the Great Supper	22:1-14		14:15-24	
The Parable of the Lost Coin			15:8-10	
The Parable of the Prodigal Son and his Brother			15:11-32	
The Parable of the Unjust Steward			16:1-9	
On Faithfulness in What is Least			16:10-12	
The Pharisees Reproved			16:14-15	
Concerning Divorce and Celibacy	19:3-12	10:2-12	16:18	
The Parable of the Rich Man and Lazarus			16:19-31	
We are Unprofitable Servants			17:7-10	
The Cleansing of the Ten Lepers			17:11-19	
On the Coming of the Kingdom of God			17:20-21	
The Parable of the Widow and Unjust Judge			18:1-8	
The Pharisee and the Publican			18:9-14	
Jesus Remains in Galilee				7:1-9
Journey to Jerusalem in Secret				7:10-13
Teaching in the Temple				7:14-39
Division among the People regarding Jesus				7:40-52
The Woman Caught in Adultery				7:53-8:11
“I am the Light of the World”				8:12-20
Discussion with the Jews				8:21-29
“The Truth will Make You Free”				8:30-36
Children of the Devil				8:37-47
“Before Abraham was, I am”				8:48-59
Jesus Heals the Man Born Blind				9:1-41
“I am the Good Shepherd”				10:1-18
Division among the Jews again				10:19-21
Jesus Blesses the Children	19:13-15	10:13-16	18:15-17	
The Rich Young Man	19:16-22	10:17-22	18:18-23	
On Riches and the Rewards of Discipleship	19:23-30	10:23-31	18:24-30 22:28-30	
The Parable of the Laborers in the Vineyard	20:1-16	10:31	13:30	
Jesus at the Feast of Dedication in Jerusalem				10:22-39
Jesus Withdraws across the Jordan				10:40-42
The Raising of Lazarus				11:1-44

The Chief Priests and Pharisees Take Counsel against Jesus				11:45-53
Jesus Retires to Ephraim				11:54-57
The Third Prediction of the Passion	20:17-19	10:32-34	18:31-34	
Jesus and the Sons of Zebedee; Precedence among the Disciples	20:20-28	10:35-45	22:24-27	
Zacchaeus			19:1-10	
The Plot against Lazarus				12:9-11
The Triumphal Entry into Jerusalem	21:1-9	11:1-11	19:28-40	12:12-19
Jesus Weeps over Jerusalem			19:41-44	
The Cursing of the Fig Tree	21:18-19	11:12-14		
The Chief Priests and Scribes Conspire against Jesus		11:18-19	19:47-48	
The Lesson from the Withered Fig Tree	21:20-22	11:20-26		
The Question about Jesus' Authority	21:23-27	11:27-33	20:1-8	
The Parable of the Two Sons	21:28-32			
The Parable of the Wicked Husbandmen	21:33-46	12:1-12	20:9-19	
The Parable of the Great Wedding Dinner	22:1-14		14:15-24	
On Paying Tribute to Caesar	22:15-22	12:13-17	20:20-26	
The Question about the Resurrection	22:23-33	12:18-27	20:27-40	
The Question about David's Son	22:41-46	12:35-37a	20:41-44	
Woe to the Scribes and Pharisees	23:1-36	12:37b-40	20:45-47	
Jesus' Lament over Jerusalem	23:37-39		13:34-35	
The Poor Widow's Gift of two Mites		12:41-44	21:1-4	
Prediction of the Destruction of the Temple	24:1-2	13:1-2	21:5-6	
Signs before the End	24:3-8	13:3-8	21:7-11	
The Desolating Sacrilege	24:15-22	13:14-20	21:20-24	
False Christs and False Prophets	24:23-28	13:21-23	17:22-24	
The Coming of the Son of Man	24:29-31	13:24-27	21:25-28	
The Time of the Coming. the Parable of the Fig Tree	24:32-36	13:28-32	21:29-36	
The Parable of the Flood and Exhortation to Watchfulness	24:37-44	13:33-37	17:26-37 12:39-40	
The Parable of the Good Servant and the Wicked Servant	24:45-51		12:41-46	
The Parable of the Ten Virgins	25:1-13			

The Parable of the Talents	25:14-30		19:11-27	
The Last Judgment	25:31-46			
The Ministry of Jesus in Jerusalem			21:37-38	
Greeks Seek Jesus; Discourse on His Death				12:20-36
The Unbelief of the People				12:37-43
Judgment by the Word				12:44-50
Jesus' Death is Premeditated	26:1-5	14:1-2	22:1-2	
The Betrayal by Judas	26:14-16	14:10-11	22:3-6	
Preparation for the Passover	26:17-20	14:12-17	22:7-14	
Washing the Disciples' Feet				13:1-20
Jesus Foretells His Betrayal	26:21-25	14:18-21	22:21-23	13:21-30
The Last Supper	26:26-29	14:22-25	22:15-20	
The New Commandment of Love				13:31-35
Peter's Denial Predicted	26:30-35	14:26-31	22:31-34	13:36-38
The Two Swords			22:35-38	
“Let Not Your Hearts be Troubled”				14:1-14
The Promise of the Paraclete				14:15-26
The Gift of Peace				14:27-31
Jesus the True Vine				15:1-8
“Abide in My Love”				15:9-17
The World's Hatred				15:18-25
The Witness of the Paraclete				15:26-27
On Persecutions				16:1-4
The Work of the Paraclete				16:5-15
Sorrow Turned to Joy				16:16-22
Prayer in the Name of Jesus				16:23-28
Prediction of the Disciples' Flight				16:29-33
The Intercessory Prayer				17:1-26
Jesus in Gethsemane	26:36-46	14:32-42	22:39-46	18:1
Jesus Arrested	26:47-56	14:43-52	22:47-53	18:2-12
Jesus before the Sanhedrin	26:57-68	14:53-65	22:54	18:13-24
Peter's Denial	26:69-75	14:66-72	22:55-62	18:25-27
Jesus Delivered to Pilate	27:1-2	15:1	23:1	18:28
The Death of Judas	27:3-10			
The Trial before Pilate	27:11-14	15:2-5	23:2-5	18:29-38
Jesus before Herod			23:6-12	
Pilate Declares Jesus Innocent			23:13-16	
Jesus or Barabbas?	27:15-23	15:6-14	23:17-23	18:39-40
Pilate Delivers Jesus to be Crucified	27:24-26	15:15	23:24-25	19:16
Jesus Mocked by the Soldiers	27:27-31a	15:16-20a		19:1-15

The Road to Golgotha	27:31b-32	15:20b-21	23:26-32	19:17
The Crucifixion	27:33-37	15:22-26	23:33-34	19:18-27
Jesus Derided on the Cross	27:38-43	15:27-32a	23:35-38	
The Two Thieves	27:44	15:32b	23:39-43	
The Death of Jesus	27:45-54	15:33-39	23:44-48	19:28-30
Witnesses of the Crucifixion	27:55-56	15:40-41	23:49	19:25-27
Jesus' Side Pierced				19:31-37
The Burial of Jesus	27:57-61	15:42-47	23:50-56	19:38-42
The Guard at the Tomb	27:62-66			
The Women at the Tomb	28:1-8	16:1-8	24:1-12	20:1-13
Jesus Appears to the Women	28:9-10	16:9-11	24:10-11	20:14-18
The Report of the Guard	28:11-15			
Jesus Appears to Two on the Way to Emmaus		16:12-13	24:13-35	
Jesus Appears to His Disciples (Thomas being Absent)			24:36-43	20:19-23
Jesus Appears to His Disciples (Thomas being Present)				20:24-31
Jesus Appears to the Eleven While They Sit at Table		16:14-18		
Jesus Appears to the Eleven on a Mountain in Galilee	28:16-20			
Jesus Appears to His Disciples by the Sea of Tiberias				21:1-25
Jesus' Ascension		16:19-20	24:44-53	

As can be easily observed from that table, there are different numbers of "items" (events) in the four gospels; there is more agreement among the three synoptic gospels (Matthew, Mark, and Luke) than there is between each of them and John; etc. But the amount of agreement needs to be quantified. I have accordingly added another table (Table 2) in which I have displayed for each pair of evangelists the percent of the time they agree with each other: number of agreements divided by the maximum number of times they could have agreed (given that there are differing numbers of items in each of the four gospels, ranging from 85 in John to 192 in Luke.)

Table 2: Between-evangelist item agreement

	Mark	Luke	John
Matthew (176items)	104/118 88.14%	139/176 78.98%	29/85 34.12%
Mark (118items)		85/118 72.34%	30/85 35.29%
Luke (192items)			29/85 34.12%
John (85items)			

As indicated in that table, and as is well-known, the strongest agreement is between Matthew and Mark. (Most biblical scholars claim that Matthew based much of his gospel on that of Mark, which is believed to have been written earlier.)

Their validity

Here are a few excerpts that provide at least partial historical corroborations of some of the gospel events. I have drawn upon Josephus as the principal source for these "confirmations".

1. Josephus, in his Antiquities of the Jews, Book 18, Chapter 3, Paragraph 3, says:

"Now there was about this time Jesus, a wise man, if it be lawful to call him a man; for he was a doer of wonderful works, a teacher of such men as receive the truth with pleasure. He drew over to him both many of the Jews and many of the Gentiles. He was [the] Christ. And when Pilate, at the suggestion of the principal men amongst us, had condemned him to the cross, those that loved him at the first did not forsake him; for he appeared to them alive again the third day; as the divine prophets had foretold these and ten thousand other wonderful things concerning him. And the tribe of Christians, so named from him, are not extinct at this day."

That paragraph, which has been regarded as authentic by some scholars and has been questioned by others, would appear to lend validity to several passages in the four gospels, namely those concerned with Jesus' miracles, His ability to attract large crowds in addition to His disciples, His being Christ, His suffering and death at the hands of Pilate, His betrayal by Judas, His denial by

Peter, and His subsequent appearances to His followers. Some of those passages are (see Table 1):

Matthew 4:18-22, 8:14-15, 8:16-17, 8:1-4, 9:1-8, 12:9-14,
4:24-25/12:15-21, 10:1-16, 8:5-13, 8:23-27, 8:28-34,
9:18-26, 9:27-31/20:29-34, 9:32-34/12:22-24, 14:13-21,
14:22-33, 14:34-36, 15:21-28, 15:29-31, 15:32-39,
17:14-18, 26:14-16, 26:47-56, 26:69-75, 27:1-2,
27:11-14, 27:24-26, 27:33-37, 27:45-54, 28:9-10, 28:16-20
[31 items]

Mark 1:16-20, 1:23-28,1:29-31, 1:32-34, 1:40-45, 2:1-12, 3:1-6, 3:7-12, 3:13-19/6:7-13, 4:35-41, 5:1-20, 5:21-43, 10:46-52,
6:32-44, 6:45-52, 6:53-56, 7:24-30, 7:31-37, 8:1-10,
8:22-26, 9:14-29, 14:10-11, 14:43-52, 14:66-72, 15:1,
15:2-5, 15:15, 15:22-26, 15:33-39, 16:9-11, 16:12-13,
16:14-18
[32 items]

Luke 4:33-37, 4:38-39, 4:40-41, 5:1-11, 5:12-16, 5:17-26, 6:6-11,
6:17-19, 6:12-16/9:1-6, 7:1-10/13:28-29, 7:11-17, 8:22-25, 8:26-39, 8:40-56, 18:35-43, 11:14-15, 9:10b-17, 9:37-43a, 13:10-17, 14:1-6, 17:11-19, 22:3-6, 22:47-53,
22:55-62, 23:1, 23:2-5, 23:13-16, 23:24-25, 23:33-34,
23:44-48, 24:10-11, 24:13-35, 24:36-43
[33 items]

John 1:35-51, 2:1-11, 4:46b-54, 5:2-47, 6:1-15, 6:16-21, 6:22-25,
9:1-41, 11:1-44, 16:29-33, 18:25-27, 18:28, 18:29-38, 19:16, 19:18-27, 19:28-30, 20:14-18, 20:19-23, 20:24-31, 21:1-25
[20 items]

2. Some evidence regarding the validity of a few other gospel items might be found in Paragraph 1, Chapter 9, Book 20 of Josephus' Antiquities, which reads in part: "... Festus was now dead, and Albinus was but upon the road; so he assembled the sanhedrim of judges, and brought before them the brother of Jesus, who was called Christ, whose name was James, and some others, [or, some of his companions]; and when he had formed an accusation against them as breakers of the law, he delivered them to be stoned..." (Scholars also disagree about the authenticity of this writing of Josephus and whether or not Jesus had a brother.) Those items are:

Matthew 10:17-25/24:9-14, 10:37-39, 10:40-42, 12:46-50,
19:23-30, 24:23-28
[6 items]

Mark 13:9-13, 3:31-35, 10:23-31, 13:21-23

[4 items]

Luke 14:25-27, 10:16, 8:19-21, 18:24-30/22:28-30,17:22-24
[5 items]

3. Another section of Antiquities (Book 18, Chapter 5, Paragraph 2) makes some explicit references to John the Baptist. It reads:

"Now some of the Jews thought that the destruction of Herod's army came from God, and that very justly, as a punishment of what he did against John, that was called the Baptist: for Herod slew him, who was a good man, and commanded the Jews to exercise virtue, both as to righteousness towards one another, and piety towards God, and so to come to baptism; for that the washing [with water] would be acceptable to him, if they made use of it, not in order to the putting away [or the remission] of some sins [only], but for the purification of the body; supposing still that the soul was thoroughly purified beforehand by righteousness. Now when [many] others came in crowds about him, for they were very greatly moved [or pleased] by hearing his words, Herod, who feared lest the great influence John had over the people might put it into his power and inclination to raise a rebellion, (for they seemed ready to do any thing he should advise,) thought it best, by putting him to death, to prevent any mischief he might cause, and not bring himself into difficulties, by sparing a man who might make him repent of it when it would be too late. Accordingly he was sent a prisoner, out of Herod's suspicious temper, to Macherus, the castle I before mentioned, and was there put to death. Now the Jews had an opinion that the destruction of this army was sent as a punishment upon Herod, and a mark of God's displeasure to him."

At the very least, that section supports the validity of Matthew 3:1-6, 3:7-10, 3:11-12, and 14:3-12; Mark 1:2-6, 1:7-8, and 6:17-29; Luke 3:1-6, 3:7-9, 3:10-14, and 3:19-20; and John 1:19-23 and 1:24-28; i.e., an additional four items for Matthew, three for Mark, four for Luke, and two for John.

4. There are a few other extra-biblical sources that might provide some evidence for the validity of other gospel items. For example, the existence of a Roman census at the time supports Luke 2:1-7. And if the Shroud of Turin is the burial cloth in which Jesus was wrapped, Matthew 27:57-61; Mark 15:42-47; Luke 23:50-56; and John 19:38-42 might also be valid. That is one more item for Matthew, one for Mark, two for Luke, and one for John.

5. Members of the controversial "Jesus Seminar" met several times about 20 years ago and tried to reach some consensus regarding the extent to which various sayings and deeds could be attributed to Jesus. The results of those meetings were compiled in two long books, the first concerned with what Jesus said (Funk, et al., 1993; they also included the gnostic gospel of Thomas) and the second concerned with what Jesus did (Funk, et al., 1998; they also included

the gnostic gospel of Peter). The methodology they employed was a type of statistical approach whereby each saying and each deed was given a rating by each seminar member on the following four-point scale by dropping a colored bead into a voting box:

A red bead indicated the voter believed Jesus did say the words in the passage quoted (did the deed described), or something very much like it. (3 Points)

A pink bead indicated the voter believed Jesus probably said (did) it. (2 Points)

A grey bead indicated the voter believed Jesus probably did not say (do) it. (1 Point)

A black bead indicated the voter believed Jesus did not say (do) it, and that it came from later admirers or a different tradition. (0 Points)

An average was taken across the ratings and subsequently converted to a proportion of maximum possible, a rank order, and a "consensus" red, pink, grey, or black. For example, the "Turn the other cheek" passage that appears in both Matthew 5:39 and Luke 6:29b received the highest rating (proportion .92, rank 1, red), whereas the "Saving one's life" passage that appears in all four gospels received a variety of ratings, ranging from a consensus black for Mark 8:35 to a consensus pink for Luke 17:33. (See pp. 549-553 in Funk, et al., 1993.)

Do those ratings add any additional items to the validity list? I've looked at the data in both of the Funk et al. books and found support (consensus red) for the following items that I have not previously indicated above:

Matthew 5:38-42, 5:43-48, 13:33, 22:15-22, 20:1-16, 6:7-15,
3:13-17, 9:10-13, 1:18-25

Mark 12:13-17, 1:9-11, 1:14-15, 2:15-17

Luke 6:29-30, 6:20-23, 6:27-28/32-36; 13:20-21, 20:20-26,
10:29-37, 16:1-9, 11:1-4, 3:21-22

If I have counted correctly, the total numbers of defensibly valid items in the four gospels are

Matthew: 53 out of a possible 176, for a "validity index" of 30.11%
Mark: 44 out of 118, validity index = 37.29%
Luke: 53 out of 192, validity index = 27.60%
John: 23 out of 85, validity index = 27.06%

Those numbers are smaller than many scholars would claim, larger than most skeptics would accept, and suggest that Mark is "the most valid". But extreme caution must be observed in interpreting such data.

Questions and comments

As food for thought, I would like to devote the remainder of this paper to some questions and comments (in no particular order) regarding the information contained in Tables 1-2 and in the previous section. In the process of so doing, I would like to call additional attention to the similarities and the differences between the four gospels and four forms of a standardized test.

1. The parallelism of the four gospels is OK but not great: quite different numbers of items, low correlations of John with the other three. In educational testing a small number of items usually results in lower reliability as well. But in fairness to the evangelists, they didn't have the time, the money, or the interest in making their gospels parallel that the Educational Testing Service has!

2. Why are some of the gospel items so short and some so long? A few of the items that appear in all four gospels even vary considerably in length, e.g., "The death of Jesus", to which Matthew devotes 10 verses, Mark 7, Luke 5, and John 13. That sort of thing happens occasionally in standardized testing; it's not necessary that parallel items have exactly the same numbers of words.

3. Table 1 combines a few pairs of events that might be referred to separately, e.g., "Jairus' Daughter and the Woman with a Hemorrhage", although both are intertwined in each of Matthew, Mark, and Luke. In standardized testing that item would be called "double-barreled". Such items are generally frowned upon, since a person might be knowledgeable about part of the item but unfamiliar with the other part.

4. On the other hand, there are a few instances in which more than one reference is made to the same event. For example, both Mark and Luke make two references to "The Commissioning of the Twelve Apostles". Duplicate or near-duplicate items are also frowned upon in standardized testing.

5. Are Bethesda and Bethsaida the same town? I assume so. Are the Gadarene demoniacs and the Gerasene demoniacs the same persons? I assume that also. Differences in the wordings of place names are quite common in the New Testament. That's not acceptable in standardized testing.

6. What about the various people named John or James? Apparently there are at least three Johns (the Baptist, the apostle, the evangelist) and at least three Jameses (two of the apostles and Jesus' alleged brother). That would wreak havoc on standardized tests.

7. Would different versions of the bible have differing reliability and validity? That matter is similar to having different versions of college aptitude tests, e.g., the SAT and its longtime competitor, the ACT (American College Testing) program .

8. The numbers of men from whom demons were cast out and the numbers of blind men whose sight was restored occasionally varies from gospel to gospel (usually one vs. two) despite the appearance of those events in two or more of the gospels, thus decreasing their parallelism.

9. John the Baptist's beheading appears in both Matthew and Mark, but much earlier chronologically in Mark. That sort of thing is not a problem in standardized testing, since the items often appear in random order, so that "What are the roots of the following quadratic equation...?" might be one of the first few items on Form A of the quantitative portion of the SAT and one of the last few on Form B.

10. Mark was found to be one of the most reliable. It was also found to be the most valid. Is that always the case with standardized tests? Why or why not?

After reading this paper, how would you answer the following questions (on a scale of 1 to 10, with 1=very bad and 10=very good)?

(1) How reliable are the four gospels?
(2) How valid are the four gospels?

APPENDIX F: The Reliability and validity of the claim that secondhand smoke causes 3,000 lung cancer deaths each year in the U.S.

Introduction

"Secondhand smoke causes approximately 3,000 lung cancer deaths among U.S. nonsmokers each year." (OSH/CDC, 2006) Do people agree about that? Is it true? In other words, is it a reliable claim? Is it valid? In what follows I shall attempt to determine the extent to which that claim regarding cause of death is consistent from one source to another and is accurate. But first some important terms must be defined and/or explicated.

The distinction between reliability and validity

In order for a claim or a measurement to be reliable it must be consistent from claimant to claimant or from measuring instrument to measuring instrument, whether or not it is accurate. In order for a claim or a measurement to be valid it must be declared to be accurate by reference to a "gold standard" of some sort. For example, a claim such as "the probability of two heads in four tosses of a fair coin is 2/4, or 1/2, or .5" is highly reliable (lots of people think that is the right answer), but it is invalid (the correct answer is 6/16= 3/8=.375; do the math). Ideally, we would like any claim to be both reliable and valid. [For more on reliability and validity see Knapp (1985) or any good measurement text such as Dunn (2004) or Borsboom (2005).]

Some preliminary remarks regarding causality in general

The usual criteria for the defensibility of a claim that something, X, causes something else, Y, are: (1) association, i.e., there is a strong relationship between X and Y; (2) temporal precedence, i.e., X comes before Y; and (3) non-spuriousness, i.e., the relationship between X and Y does not vanish when other variables U,V, W, etc. are taken into account. Some authorities specify additional criteria such as (4) theoretical meaningfulness, i.e., it makes scientific sense; and/or (5) "dose response", i.e., the greater the X the much greater the Y. Those criteria are more likely to be satisfied by true experiments (randomized clinical trials) than by observational research.

For the secondhand smoke claim, if there is a strong relationship between amount of exposure to secondhand smoke (X) and mortality from lung cancer (Y, a died vs. lived dichotomy); if the exposure occurred before the mortality (in this case, how else?!); if there is no confounding factor (e.g., concurrent exposure to some other possible cause that might account for the relationship); if the claim is biologically plausible; and if there is empirical evidence for a dose response; then the allegation "secondhand smoke causes lung cancer deaths" would be supported. How many such deaths there are each year is a separate matter.

Much has been written about causality in the literature of philosophy, epidemiology, statistics, and other scientific disciplines. My favorite sources are the articles by Holland (1986; 1993), the book by Pearl (2000), and the articles by Hernan (2004) and by Hernan and Robins (2006). All of those sources are rigorous, occasionally obtuse, and, if you skip around a bit, enlightening. Especially enlightening (and understandable) are Sections 3 and 7 in Holland's 1986 article; Sections 1.2.1, 3.3.3, 7.3.2, and the Epilogue in Pearl's book; and the delightful hypothetical data set in the Hernan articles. Section 3 in Holland (1986) is particularly relevant for individual vs. group causality (see below), as is the first part of Hernan (2004). The Epilogue in Pearl is a marvelous historical summary of the development of the concept of causality, complete with great illustrations. And Hernan's choices of Greek names (Zeus, Hera, etc.) for patients are hilarious.

A cause vs. the cause

It is important to differentiate between a claim of the form "X is a cause of Y" and a claim of the form "X is the cause of Y". It is difficult enough to obtain evidence concerning the former; it is almost impossible to obtain evidence concerning the latter. If X is a cause of Y, the occurrence of X is sufficient for Y to occur. If Y cannot occur without X, the occurrence of X is necessary for Y to occur. If X is the (one and only) cause of Y, the occurrence of X is both necessary and sufficient for Y to occur. Many causal claims in the field of public health are of the more general form "X causes Y" (e.g., "smoking causes lung cancer"), and it is therefore much more difficult to determine their reliability and their validity. (See Spirtes, Glymour, & Scheines, 2000, esp. pp. 239-249, for an excellent discussion of the difficulty of actually testing the hypothesis that [firsthand, mainstream] smoking causes lung cancer.)

Deterministic (structural) causality vs. probabilistic causality

A claim such as "smoking causes lung cancer" is probabilistic; i.e., "if you smoke you might get lung cancer", not "if you smoke you must get lung cancer", since it is known that there are some smokers who get lung cancer and there are some who don't; there are some nonsmokers who get lung cancer and some who don't. All we can say from the empirical evidence is that if you smoke there is a greater risk (higher probability) of getting lung cancer than if you don't smoke.

There are very few examples of deterministic causality. One that comes immediately to mind is the claim that the bullet fired from Jack Ruby's gun caused Lee Harvey Oswald's death. Many of us saw that on national television. (The claim that the bullet fired from Oswald's gun was the cause of President Kennedy's death is extremely controversial, albeit deterministic, and in the eyes of conspiratorial theorists it is not even highly probable.)

An interesting concept in the legal literature is "the probability of causation" (Parascandola, 1998; Robins, 2004; Scheines, 2008; and Swaen & vanAmelsvoort, 2009), which is often used to assess liability and damages in lawsuits, especially toxic tort cases, that are brought against parties alleged by the plaintiffs to have caused injury or death. It is an attempt to reconcile individual causality with group causality, and is a function of relative risk. (See the following sections.)

Individual (singular, token, single-event) causality vs. group (general) causality

Is it possible that X caused Y for a group, but we are unable to identify any individual within that group for whom X caused Y? Apparently (see Holland, 1986; Hernan, 2004), and that is reflected in the literature concerning the effects of secondhand smoke on lung cancer. Except for a couple of high-profile cases such as Dana Reeve (nonsmoking wife of Superman-portrayer Christopher Reeve) and Heather Crowe (nonsmoking Canadian anti-tobacco activist), and an unidentified asthmatic waitress in Michigan (see Stanbury, et al., 2008), I know of no other individuals for whom it has been claimed that secondhand smoke caused the lung cancer that resulted in their deaths, and even in those three cases the evidence is not clear. Yet the claim of thousands of deaths for a group of people persists--see, for example, the Surgeon General's recent report on the effects of secondhand smoke (U.S. Department of Health and Human Services, 2006). [In that report "secondhand smoking" is referred to as "involuntary smoking". The terms "passive smoke", "sidestream smoke", and "environmental tobacco smoke" are also often used as synonyms for secondhand smoke.]

Attributable risk

Most of the epidemiological claims of causes of deaths are based upon the concept of attributable risk. That concept was originally due to Levin (1953) as one of three indexes of the relationship between firsthand smoking and lung cancer (he didn't actually use the term "attributable risk"), and is defined as follows (his notation):

$$S = b(r-1)/ [b(r-1) + 1],$$

where b is the proportion of the population under consideration that is exposed (e.g., the proportion of smokers), and r is the ratio of the incidence of the disease (e.g., lung cancer) in the exposed sub-population to the incidence of the disease in the unexposed sub-population. [In the more recent epidemiological literature, S has been replaced by PAF (the population attributable fraction) or PAR (the population attributable risk), b has been replaced by p, and r has been replaced by RR (relative risk).] In order to estimate numbers of a certain kind of death attributable to an exposure, one determines the prevalence of the exposure (b)

and the relative risk of death from that exposure (r), calculates the attributable risk, and multiplies that by the total number of deaths of that kind.

["The probability of causation", PC (see above), is equal to 1 - 1/RR, and is usually required to be greater than .50, which corresponds to a relative risk of 2, in order for a judgment to be made in a plaintiff's favor.]

It is important to note that Levin claimed the attributable risk calculated by the above formula is a maximum risk (see p. 536 of his paper). It is also important to note that he assumed both b and r to have been determined for the same population (e.g., a population cohort consisting of an exposed sub-population and an unexposed sub-population). It is common practice in epidemiological research to obtain an estimate of b from one source and an estimate of r from another source, especially in case-control studies.

One of the problems with this method is that estimates of PAFs for diseases with multiple risk factors can add to more than 1 (see, for example, Rowe, Powell, and Flanders, 2004). Several explanations of, and/or alternatives to, Levin's formula for the determination of attributable risk have been suggested by Miettinen (1974), as reinforced by Hanley (2001); and by Begg, Satagopan, and Berwick (1998); Begg (2001); Eide and Heuch (2001); Uter and Pfahlberg (1999, 2001); Heller, Buchan, et al. (2003); Ha-Duong, Casman, and Morgan (2004); and others. The approach taken by Ha-Duong et al. is called "bounding analysis", and was severely criticized by Greenland (2004; see also the response by Casman, Ha-Duong, and Morgan, 2004).

Death certificates

In their article concerned with the quality of information for causes of death in 115 countries, Mathers, Fat, et al. (2005) argued that “who dies from what” is “the most basic of health statistics” (abstract). The document that is most relevant for the determination of cause of death in this country is the U.S. Standard Certificate of Death. The section on cause(s) of death (Items 31-37) contains information provided by a medical certifier. The certifier can be a coroner, a medical examiner, or a family physician. [See Magrane, Gilliland, & King (1997) and the associated editorial by Huffman (1997) for a discussion of the role of a family physician in the certification of death.]

Item 35 asks if tobacco use contributed to the death. That item had already been included in the death certificates used in Colorado, Louisiana, Maryland, Nebraska, North Dakota, Oregon, Texas, Utah, and New York City since 1989, but is new to the other states. Zevallos, Huang, et al. (2004) claimed that its addition to the Texas death certificate increased the reporting of tobacco use as a contributor to mortality. [N.B. "Tobacco use" is not the same as "exposure to secondhand smoke", but in certain situations it can be used as a surrogate for it--see below.]

Is the death certificate the gold standard for cause of death or just one source that must itself be evaluated for reliability and validity? The literature is surprisingly equivocal about that. Death certificates have been subjected to various criticisms over the years (see, for example, Gittlesohn & Royston, 1982; Feinstein, 1985; Messite & Stellman, 1996; Lenfant, Friedman, and Thom, 1998; and Lloyd-Jones, Martin, et al., 1998), but no other readily available source has been suggested that might be a more accurate indicator of cause of death. Consider, for example, the study reported by Thomas, Hedberg, and Fleming (2001). They compared the numbers of deaths estimated by epidemiologists to be attributable to cigarette smoking (using the SAMMEC software that is based upon the method for investigating group causality originally due to Levin, 1953) with a tally of deaths having cigarette smoking as the principal cause indicated on the death certificate (Item 31 on the current version) for a sample of persons who died in Oregon between 1989 and 1996. The agreement was in general quite good, but there were a few rather large discrepancies (e.g., cervical cancer: 109 for SAMMEC vs. 29 for death certificates). The authors were unwilling to claim that the death certificate was the gold standard; they regarded both approaches as alternatives that were likely to be convincing to different audiences. It is therefore best to consider their study as a reliability study.

Autopsies

It could be argued that autopsies are, or should be, the "platinum standard" for cause of death. Fortunately, or unfortunately, depending upon one's point of view, autopsies are rare, being reserved primarily for deaths that are suspected to be murders. There have been a few studies in which the validity of certain alleged causes of death has been investigated, using autopsy as the external criterion. For example, Nashelsky and Lawrence (2003) reported that medical examiners and coroners were wrong in 28% of 261 cases in determining the cause of death when blinded to actual autopsy results. And autopsies also were found to disagree with death certificates in 37% of 155 forensic exhumations (Karger, de la Grandmaison, et al., 2004).

There are also "verbal autopsies", which consist of cause-seeking interviews with the relatives and/or friends of decedents in places where there are no medical certifiers, such as Third World countries. [For further information concerning verbal autopsies see Fauveau (2006) and King & Lu (2006).]

Secondhand smoke (and numbers of lung cancer deaths)

Nothing has been more controversial and politically charged than the allegation that secondhand smoke causes thousands of deaths and those deaths are preventable by banning smoking in public places. The allegation can be written as: firsthand smoking→secondhand smoke→health problems→death. On one side of the controversy are the Surgeons General; the Centers for

Disease Control and Prevention (CDC); the American Cancer Society (ACS); the American Heart Association (AHA); the American Lung Association (ALA); and Stanton Glantz, James Repace, and other anti-smoking activists (especially New York City's Mayor Michael Bloomberg) who wholeheartedly support the claim. On the other side are the tobacco industry [naturally]; Enstrom and Kabat (2003); Elizabeth Whelan, President of the American Council on Science and Health (ACSH); and writers such as Jerry Arnett, Gio Gori, Michael McFadden, Michael Siegel, and Jacob Sullum who contend that the effects of secondhand smoke have been vastly exaggerated. There is also the critique by Feinstein (1992) of the article by Smith, Sears, et al. (1992); the opposing positions by Gross and by Rockette (1993); a subsequent article by Gross in 1995, the rejoinders by Bayard, Jinot, & Flatman and Hanley to that article, and Gross's response; the recent piece in JAMA by Kuehn (2006); the article by Stranges, Bonner, et al. (2006); and many others.

The evidence

So, what is the actual evidence regarding the reliability and the validity of the claim that "Secondhand smoke causes approximately 3,000 lung cancer deaths among U.S. nonsmokers each year"?

Reliability

The 3,000 figure has been cited in numerous sources ever since the report on the effects of secondhand smoke was issued by the U.S. Environmental Protection Agency (EPA, 1992). [The report was extremely controversial at the time, resulting in a number of lawsuits, judicial decisions, and smoking bans of various sorts. It still is controversial.] The agreement among most sources is quite good. For example, although they do not all differentiate between never smokers and former smokers (current smokers are presumed to die from their own firsthand smoke), and they are not completely independent, most of the entries returned when you Google the words "lung cancer deaths secondhand smoke") claim 3,000 to 3,400 deaths per year. The modifiers "about", "approximately", "at least", or "more than" are sometimes added, as are additional words such as "adult", "excess", "premature", and "preventable". [For a critique of the concept of "premature death", see Trisel, 2007.] The 3,400 figure comes from an updated analysis carried out by the CalEPA (2005) and is cited by both the American Cancer Society and the American Lung Association on their websites. [At one point in its report, on page 7-62, CalEPA claims "a range of 3423 to 8866".] The only claims I could find that were considerably different from 3,000 (other than those who argue that there are at most a handful) were an estimate of 5,000 (actually 4,665 rounded to 5,000) made by Repace and Lowrey (1985) and an estimate of 300,000 attributed to former surgeon general Dr. David Satcher in his 1999 G. Gayle Stephens lecture (Coastal Research Group, 2004), but the latter might very well have been a

"voiceographical" or typographical error. There is therefore reasonably strong evidence for the reliability of the claim.

Some observations regarding internal consistency:

1. The American Cancer Society estimated that in 2007 there would be approximately 160,000 lung cancer deaths (ACS, 2007), 85-90% of which would be to smokers (Thun, 2006) and the remaining 10-15% to nonsmokers, giving an estimate of approximately 20,000 lung cancer deaths for nonsmokers. ACS has further estimated that there are approximately 11,000 lung cancer deaths per year for never smokers (a subset of nonsmokers), 27% of which are attributable to secondhand smoke (ACS, 2006), i.e., approximately 2,970 such deaths. That is a bit lower than their 3,400 estimate for all nonsmokers (the other 430 are former smokers?).

2. It is interesting to analyze the data for the two states that have the highest (Kentucky) and the lowest (Utah) prevalences of cigarette smoking. According to the 2000 Census (see Table 5 in its December 28, 2000 release), the population of Kentucky was 4,041,769, which was approximately 1.436% of the total U.S. resident population of 281,421,906. If the number of lung cancer deaths for nonsmokers in the country as a whole were 3,000, and if Kentucky had a proportionate share, then there would have been approximately 43 such deaths in that state. The prevalence of smoking in Kentucky in the year 2000 was estimated to be 30.5% (MMWR, 2001). The relative risk of lung cancer for those exposed to secondhand smoke vs. those not exposed has been variously reported as ranging between 1.20 and 1.30 (i.e., an elevated risk of 20 to 30%), although Enstrom and Kabat (2003) estimated it to be very close to 1.00. The prevalence of firsthand smoking provides an estimate of the probability of exposure of nonsmokers to the smoke emitted by smokers (see page 8-38 of the CalEPA report for an example of an analysis of the number of cardiovascular deaths from secondhand smoke based upon that assumption). Using that estimate of a p of .305 together with an RR of 1.25 in the SAMMEC version of Levin's attributable risk formula, a PAF of .0708 is obtained. The age-adjusted lung cancer death rate in Kentucky in 2000 was estimated to be 80.2 per 100,000 persons. Applying that estimate to Kentucky's population in that year yields 3,242 lung cancer deaths. .0708 (3,242) = approximately 230 of them would be estimated to be for nonsmokers. The 43 and the 230 are reconcilable if the 43 is for never smokers and the 230 is for all nonsmokers; i.e., the difference of 187 is for former smokers. As indicated above, many of the claimants do not clarify whether the estimate is for never smokers only or for never smokers and former smokers combined.

Utah's population in 2000 was reported to be 2,233,169, or approximately .793% of the total U.S. resident population. Its proportionate share of 3,000 lung cancer deaths for nonsmokers would be approximately 24. Utah's smoking prevalence was 12.9% (MMWR, 2001). For a p of .129 and an RR of 1.25, PAF = .0312.

The age-adjusted lung cancer death rate in Utah was estimated to be 26.3 per 100,000, yielding 587 lung cancer deaths, which when multiplied by .0312 is 18 for nonsmokers. The 24 and the 18 are not reconcilable. If the former number is for never smokers only, the latter number does not make sense, since subtraction of the 24 from the 18 would yield a figure of -6 lung cancer deaths for former smokers.

The 43 for Kentucky and the 24 for Utah are commensurate with what would be expected for their respective populations in conjunction with the ACS's estimate of 2.970 lung cancer deaths for never smokers attributable to secondhand smoke in the entire U.S. per year.

Validity

Unfortunately there is no consensus regarding what constitutes a gold standard for the accuracy of the claim. The closest thing is the analysis carried out and reported by the EPA in 1992, which, as indicated above, was the subject of considerable controversy. In my opinion it was actually a laudable attempt (see Chapter 6 of that report) to estimate the annual number of lung cancer deaths (particularly for nonsmoking adult females) that might have been attributable to secondhand smoke up until that time (emphasis mine) and it has been frequently cited as the basis for the 3,000 figure. The actual EPA estimate for the year 1985 was 3,060 and included both never smokers (2000 deaths) and former smokers (1060 deaths);1130 male deaths, 1930 female deaths; 860 deaths from exposure at home, 2200 deaths from exposure at work--see their Table 6-3). The amazing thing, however, is that the claim has essentially remained fixed at 3,000 ever since, despite the fact that there has been a decrease in firsthand smoking, and consequently a decrease in exposure to secondhand smoke, in the intervening years (see, for example, Soliman, Pollack, & Warner, 2004; Pirkle, Bernert, et al., 2006; CDC, 2008). The determination of the accuracy of the claim is admittedly extremely difficult. Not only is it almost impossible to agree upon a gold standard, but there are additional problems:

1. The definition of smokers and nonsmokers. [Smokers are almost always equated with cigarette smokers, with little or no attention to pipe smokers or cigar smokers.] In most of the literature (see, for example, MMWR, 2001) smokers are defined as persons who have smoked at least 100 cigarettes in their lifetimes and are currently smoking every day or on some days. Nonsmokers are divided into former smokers and never smokers. Former smokers are said to have smoked at least 100 cigarettes in their lifetimes but are not currently smoking. Never smokers are those who have smoked fewer than 100 cigarettes in their lifetimes. Such definitions are both vague and debatable.

2. The measurement of amount of exposure to secondhand smoke. The most common approach is the use of questionnaires, in which respondents are asked to recall how much they have been exposed to secondhand smoke in their

homes and/or at their workplaces. Biomarkers such as urine cotinine, saliva cotinine, or hair nicotine have also been suggested--see, for example, Benowitz (1996), Al-Delaimy, Crane, & Woodward (2002), Siegel and Skeer (2003), Stark, Rohde, et al. (2007), and Okoli, Hall, et al. (2007)--as have airborne particle monitors for outdoor tobacco smoke (Klepeis, Ott, & Switzer, 2007).

3. The need for constant updating of the prevalence of smoking, the relative risk of death for the exposed vs. the unexposed, and the numbers of lung cancer deaths.

Cardiovascular deaths

Reference was made above to an analysis of the estimated number of cardiovascular deaths that are attributable to secondhand smoke (CalEPA, 2005, page 8-38). That estimate was a range of 22,669 to 69,553 deaths, with a midpoint of 46,111. (See also Glantz & Parmley, 1991, 1995, and 2001 for similar claims.) Those claims are considerably less reliable and their validity is equally problematic. (See, for example, Nilsson, 2001 and Enstrom & Kabat, 2006.) Mayor Bloomberg once claimed that secondhand smoke caused 1,000 deaths each year in New York City alone (a figure that Elizabeth Whelan argued was much too high). The city's proportionate share of 3,000 lung cancer deaths for non- smokers for the country as a whole, given its population of approximately 8 million and a population of 300 million for the entire U.S., would be about 80. Its proportionate share of 46,000 cardiovascular deaths would be over 1,200.

Thirdhand smoke

Believe it or not, some researchers are claiming that there are effects of thirdhand smoke [and even fourthhand and fifthhand smoke] on non-smokers. The term "thirdhand smoke" was first coined by Gerald Nachman (1991) [as far as I have been able to determine], but it was recently redefined in an article by Winickoff et al. (2009) regarding the need for banning smoking in homes.

Summary

Since autopsies are rare, death certificates have a number of shortcomings, and the public health community does not seem to be able to agree upon a gold standard for cause of death, the validity of the allegation regarding the number of lung cancer deaths for nonsmokers caused by secondhand smoke is likely to remain indeterminate. The reliability of such claims will continue to be addressed, but even there the choice of sources to compare with one another presents a serious challenge.

REFERENCES

[Notes: 1. For sources that have more than three authors I have indicated only the first two, followed by et al. 2. The numbers and letters within the parentheses at the ends of the references are the designations of the chapters or appendices in which the references are cited.]

Abelson, R.P. (1995). Statistics as principled argument. Hillsdale, NJ: Erlbaum. (11)

Abelson, R.P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), What if there were no significance tests? (Chapter 5, pp. 117-141). Mahwah, NJ: Erlbaum. (11)

Ackerson, L. (1933). In disagreement with E.A. Lincoln's article, "The unreliability of reliability coefficients". Journal of Educational Psychology, 24, 233-235. (10)

ACS (2006). Causes of lung cancer in nonsmokers. (Stat Bite). Journal of the National Cancer Institute, 98 (10), 664. (F)

ACS (2007). Cancer facts and figures, 2007. Atlanta, GA: American Cancer Society. (F)

Adams, H.F. (1936). Validity, reliability, and objectivity. Psychological Monographs, 47, 329-350. (2)

Agresti, A. (1984). Analysis of ordinal categorical data. New York: Wiley. (10)

Agresti, A., & Finlay, B. (1997). Statistical methods for the social sciences (3rd. ed.). Upper Saddle River, NJ: Prentice-Hall. (3)

Aiken, L.R. (1966). Another look at weighting test items. Journal of Educational Measurement, 3, 183-185. (13)

Aiken, L.R. (1988). A program for computing the reliability and maximum reliability of a weighted composite. Educational and Psychological Measurement, 48, 703-706. (13)

Aiken, L.S., & West, S.G. (1991). Multiple regression: Testing and interpreting interactions. Newbury Park, CA: Sage. (11)

Al-Delaimy, W.K., Crane, J., & Woodward, A. (2002). Is hair nicotine level a more accurate biomarker of environmental tobacco smoke exposure than urine cotinine? Journal of Epidemiology and Community Health, 56, 66-71. (F)

Altman, D.G., & Bland, J.M. (1983). Measurement in medicine: The analysis of method comparison studies. Statistician, 32, 307-317. (3)

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association. (1)

Armor, D. (1974). Theta reliability and factor scaling. In H. Costner (Ed.), Sociological Methodology 1974 (pp. 17-50). San Francisco: Jossey-Bass. (8)

Armstrong, G.D. (1981). The intraclass correlation as a measure of interrater reliability of subjective judgments. Nursing Research, 30, 314-315, 320A. (9)

Bacon, D. (2004). The contributions of reliability and pretests to effective assessment. Practical Assessment, Research & Evaluation, 9 (3).

Baker, S.G., & Kramer, B.S. (2003). A perfect correlate does not a surrogate make. BMC Biomedical Research Methodology, 3 (16), 1-5. (10,D)

Barnette, J.J. (2005). ScoreRel CI: An Excel program for computing confidence intervals for commonly used score reliability coefficients. Educational and Psychological Measurement, 65 (6), 980-983. (11)

Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19, 3-11. (9)

Bartko, J.J. (1976). On various intraclass correlation reliability coefficients. Psychological Bulletin, 83, 762-765. (9)

Bartko, J.J., & Carpenter, W.T. (1976). On the methods and theory of reliability. The Journal of Nervous and Mental Disease, 163, 307-317. (9)

Bashir, S.A., & Duffy, S.W. (1997). The correction of risk estimates for measurement error. Annals of Epidemiology, 7, 154-164. (2)

Bashir, S.A., Duffy, S.W., & Qizilbach, N. (1997). Repeat measurements of case-control data: Corrections for measurement error in a study of ischaemic stroke and haemostatic factors. International Journal of Epidemiology, 26, 64-70. (2)

Bayard, S., Jinot, J, & Flatman, G. Environmental tobacco smoke and lung cancer: Uncertainties in the population estimates but not in the causal association - A rejoinder to Gross. Environmetrics, 6 (4), 413-418. (F)

Bechtoldt, H.P. (1963). Correlational methods in research on human learning: An amplification. Perceptual and Motor Skills, 16, 831-842. (6)

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. Psychological Methods, 5 (3), 370-379. (2)

Begg, C.B. (2001). The search for cancer risk factors: When can we stop looking? American Journal of Public Health, 91 (3), 360-364. (F)

Begg, C.B., Satagapon, J.M. & Berwick, M. (1998). A new strategy for evaluating the impact of epidemiologic risk factors for cancer, with application to melanoma. Journal of the American Statistical Association, 93 (442), 415-426. (F)

Benowitz, N.L. (1996). Cotinine as a biomarker of environmental tobacco smoke exposure. Epidemiological Review, 18, 188-204. (F)

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C.W. Harris (Ed.), Problems in measuring change (Chapter 1, pp. 3-20). Madison, WI: The University of Wisconsin Press. (6)

Berk, R.A. (1980). A consumer's guide to criterion referenced reliability. Journal of Educational Measurement, 17, 323-349. (13)

Berk, R.A. (2000). Ask Mr. Assessment Person: How do you estimate the reliability of teacher licensure/certification tests? In Teachers: Supply and demand in an age of rising standards. National Evaluation Systems, Inc. (3)

Beyer, J.E., Turner, S.B., et al. (2005). The alternate forms reliability of the Oucher Pain Scale. Pain Management Nursing, 6 (1), 10-17. (1)

Biswas, A.K. (2006). Reliability of total test scores when considered as ordinal measurements. Applied Psychological Measurement, 30 (1), 43-55. (10)

Blalock, H.M., Jr., & Blalock, A.B. (Eds.) (1968). Methodology in social research. New York: McGraw-Hill. (6)

Blanchard, B.S. (1981). Logistics and engineering management. Englewood, NJ: Prentice-Hall. (1)

Bland, J.M., & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 1 (8476), 307-310. (3,6)

Bland, J.M., & Altman, D.G. (1999). Measuring agreement in method comparison studies. Statistical Methods in Medical Research, 8, 135-160. (3)

Bobko, P. (1983). An analysis of correlations corrected for attenuation and range restriction. Journal of Applied Psychology, 68, 584-589. (4)

Bock, R.D. (1997). A brief history of item response theory. Educational Measurement: Issues and Practice, 15 (4), 21-33. (13)

Bock, R.D., & Wood, R. (1971). Test theory. Annual Review of Psychology, 22, 193-224. (3)

Bohrnstedt, G.W. (1969). Observations on the measurement of change. In E.F. Borgatta (Ed.), Sociological Methodology 1969 (pp. 113-136). San Francisco: Jossey-Bass. (6)

Bohrnstedt, G.W. (1983). Measurement. In P.H. Rossi, J.D. Wright, & A.B. Anderson (Eds.), Handbook of survey research (Chapter 3, pp. 69-121). New York: Academic Press. (3,4,7)

Bollen, K.A. (1989). Structural equation modeling with latent variables. New York: Wiley. (3,13)

Bollen, K.A. (2002). Latent variables in psychology and the social sciences. Annual Review of Psychology, 53, 605-634. (3)

Bond, L. (1979). On the base-free measure of change proposed by Tucker, Damarin, and Messick. Psychometrika, 44, 351-355. (6)

Bonett, D.G. (2002). Sample size requirements for testing and estimating coefficient alpha. Journal of Educational and Behavioral Statistics, 27 (4), 335-340. (11)

Bonett, D.G., & Wright, T.A. (2000). Sample size requirements for estimating Pearson, Spearman, and Kendall correlations. Psychometrika, 65, 23-28. (11)

Borman, W.C., Buck, D.E., et al. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. Journal of Applied Psychology, 86 (5), 965-983. (1)

Borsboom, D. (2005). Measuring the mind: Conceptual issues in contemporary psychometrics. Cambridge (England): Cambridge University Press. (2, B)

Borsboom, D., & Mellenbergh, G.J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. Intelligence, 30, 505-514. (2)

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2002). The concept of validity. Psychological Review, 111, 1061-1071. (B)

Bradley, J.L., Brown, J.E., & Himes, J.H. (2001). Reliability of parental measurements of infant size. American Journal of Human Biology, 13, 275-279. (D)

Braun, H.I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. Journal of Educational Statistics, 13 (1), 1-18. (1)

Bravo, G., & Potvin, L. (1991). Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: Toward the integration of two traditions. Journal of Clinical Epidemiology, 44 (4/5), 381-390. (9)

Brener, N.D., McManus, T., et al. (2003). Reliability and validity of self-reported height and weight among high school students. Journal of Adolescent Health, 32, 281-287. (D)

Brennan, P.F., & Hays, B.J. (1992). The kappa statistic for establishing interrater reliability in the secondary analysis of qualitative clinical data. Research in Nursing & Health, 15, 153-158. (7)

Brennan, R.L. (1975). The calculation of reliability from a split-plot factorial design. Educational and Psychological Measurement, 35, 779-788. (13)

Brennan, R.L. (1997). A perspective on the history of generalizability theory. Educational Measurement: Issues and Practice, 16 (4), 14-20. (13)

Brennan, R.L. (1998). Misconceptions at the intersection of measurement theory and practice. Educational Measurement: Issues and Practice, 17 (1), 5-9, 30. (13)

Brennan, R.L. (2000). (Mis)conceptions about generalizability theory. Educational Measurement: Issues and Practice, 19 (1), 5-10. (13)

Brennan, R.L. (2001a). Generalizability theory . New York: Springer. (13)

Brennan, R.L. (2001b). An essay on the history and future of reliability from the perspective of replications. Journal of Educational Measurement, 38 (4), 295-317. (13)

Brennan, R.L., & Light, R.J. (1974). Measuring agreement when two observers classify people into categories not defined in advance. British Journal of Mathematical and Statistical Psychology, 27, 154-163. (7)

Brennan, R.L., & Prediger, D.J. (1981). Coefficient kappa: Some uses. Educational and Psychological Measurement, 41, 687-699. (7)

Brook, R.J., & Stirling, W.D. (1984). Agreement between observers when categories are not specified. British Journal of Mathematical and Statistical Psychology, 37, 271-282. (7)

Brown, J., Krieger, N., et al. (2001). Misclassification of exposure: Coffee as a surrogate for caffeine intake. American Journal of Epidemiology, 153, 815-820.

Brown, J.K., Feng, J-Y., & Knapp, T.R. (2002). Is self-reported height or arm span a more accurate alternative measure of height? Clinical Nursing Research, 11 (4), 417-432. (12, D)

Brown, J.K., Whittemore, K.T., & Knapp, T.R. (2000). Is arm span an accurate measure of height in young and middle aged adults? Clinical Nursing Research, 9, 84-94. (12, D)

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296-322. (4,8)

Brown, W. (1913). The effects of "observational errors" and other factors upon correlation coefficients in psychology. British Journal of Psychology, 6, 223-235. (4)

Brownell, W.A. (1933). On the accuracy with which reliability may be measured by correlating test halves. Journal of Experimental Education, 1, 204-215. (8)

Bruton, A., Conway, J.H., & Holgate, S.T. (2000). Reliability: What is it and how is it measured? Physiotherapy, 86 (2), 94-99. (10)

Burns, K.J. (1998). Beyond classical reliability: Using generalizability theory to assess dependability. Research in Nursing & Health, 21, 83-90. (13)

Burt, C. (1955). Test reliability estimated by analysis of variance. British Journal of Statistical Psychology, 8, Part 2, 103-118. (8)

Byrne, M. W., & Lenz, E.R. (2002). Reliability of transportable instruments for assessment of infant length. Journal of Nursing Measurement, 10 (2), 111-121. (D)

Byrt, T., Bishop, J., & Carlin, J.B. (1993). Bias, prevalence, and kappa. Journal of Clinical Epidemiology, 46 (5), 423-429. (7)

Cable, L.W. (no date) Are the gospels true? Retrieved from the author's website (www.inu.net/skeptic/) on February 18, 2009. (E)

CalEPA (2005). Proposed identification of environmental tobacco smoke as a toxic air contaminant. California Environmental Protection Agency. (F)

Callender, J.C., & Osburn, H.G. (1979). An empirical comparison of coefficient alpha, Guttman's Lambda-2, and maximized split-half reliability estimates. Journal of Educational Measurement, 16, 89-99. (8)

Cameron, J. M. (1982). Error analysis. In S. Kotz, & N. L. Johnson (Eds. In Chief), & C. B. Read (Assoc. Ed.), Encyclopedia of statistical sciences (vol. 2). Toronto: Wiley. Pp. 545-551. (2)

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multi-trait multi-method matrix. Psychological Bulletin, 56, 81-105. (B)

Campbell, D.T., & Stanley, J.C. (1966). Experimental and quasi-experimental designs for research. Chicago: Rand McNally. (B)

Carmines, E.G., & Zeller, R.A. (1979). Reliability and validity assessment. Volume 17, Quantitative Applications in the Social Sciences. Beverly Hills, CA: Sage. (8)

Carroll, R.J. (1997). Surprising effects of measurement error on an aggregate data estimator. Biometrika, 84 (1), 231-234. (4)

Casman, E.A., Ha-Duong,M., & Morgan, M.G. (2004). Response to Sander Greenland's critique of bounding analysis. Risk Analysis, 24 (5), 1093-1095. (F)

Cattell, R.B. (1964). Validity and reliability: A proposed more basic set of concepts. Journal of Educational Psychology, 55, 1-22. (2)

Cattell, R.B. (1982). The clinical use of difference scores: Some psychometric problems. Experimental Clinical Research, 6, 87-98. (6)

CDC (November 14, 2008). Cigarette smoking among adults---United States, 2007. Morbidity and Mortality Weekly Report, 57 (45), 1221-1226. (F)

Charles, E.P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. Psychological Methods, 10 (2), 206-226. (4,11).

Charter, R.A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. Journal of Clinical and Experimental Neuropsychology, 21 (4), 559-566. (11)

Charter, R.A. (2001). It is time to bury the Spearman-Brown "prophecy" formula for some common applications. Educational and Psychological Measurement, 61, 690-696. (8)

Charter, R.A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. The Journal of General Psychology, 130 (3), 290-304. (3)

Chase, C. (1996). Estimating the reliability of criterion-referenced tests before administration. Mid-Western Educational Researcher, 9 (2), 2-4. (13)

Cicchetti, D.V., & Feinstein, A.R. (1990). High agreement but low kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology, 43 (6), 551-558. (7)

Cleary, T.A., & Linn, R.L. (1969a). A note on the relative sizes of the standard errors of two reliability estimates. Journal of Educational Measurement, 6, 25-27. (11)

Cleary, T.A., & Linn, R.L. (1969b). Error of measurement and the power of a statistical test. British Journal of Mathematical and Statistical Psychology, 22 (Part 1), 49-55. (11)

Cliff, N. (1979). Test theory without true scores. Psychometrika, 44, 373-393. (2)

Cliff, N. (1984). An improved internal consistency reliability estimate. Journal of Educational Statistics, 9, 151-161. (8)

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. Psychological Bulletin, 103, 276-279. (8)

Cliff, N. (1989). Ordinal consistency and ordinal true scores. Psychometrika, 54, 75-91. (10)

Cliff, N., & Caruso, J.C. (1998). Reliable component analysis through maximizing composite reliability. Psychological Methods, 3, 291-308. (8,13)

Cliff, N., & Keats, J.A. (2003). Ordinal measurement in the behavioral sciences. Mahwah, NJ: Erlbaum. (10)

Cline, M.E., Herman, J., et al. (1992). Standardization of the visual analogue scale. Nursing Research, 41, 378-380. (7)

Coastal Research Group (2004). The Ninth G. Gayle Stephens Lecture by Surgeon General David Satcher, M.D., Ph.D. Retrievable at www.coastalresearch.org. (F)

Cochran, W.G. (1968). Errors of measurement in statistics. Technometrics, 10, 637-666. (2,4)

Coffman, W.E. (1972). On the reliability of ratings of essay examinations. Measurement in Education, 3 (3), 1-7. (1)

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46. (7,11)

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scale disagreement or partial credit. Psychological Bulletin, 70, 213-220. (7)

Cohen, J. (1983). The cost of dichotomization. Applied Psychological Measurement, 7, 249-253. (13)

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd. ed.). Hillsdale, NJ: Erlbaum. (11)

Conger, A.J. (1974). Estimating profile reliability and maximally reliable composites. Multivariate Behavioral Research, 9, 84-104. (13)

Conger, A.J. (1980). Maximal reliability composites for unidimensional measures. Educational and Psychological Measurement, 40, 367-371. (13)

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

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78, 98-104. (8)

Cronbach, L.J. (1947). Test “reliability”: Its meaning and determination. Psychometrika, 12, 1-16. (1,3)

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334. (8)

Cronbach, L.J. (1970). Essentials of psychological testing (3rd ed.). New York: Harper & Row. (2,5)

Cronbach, L.J. (1988). Internal consistency of tests: Analyses old and new. Psychometrika, 53, 63-70. (8)

Cronbach, L.J., & Furby, L. (1970). How should we measure "change"--or should we? Psychological Bulletin, 74, 68-80. (6)

Cronbach, L.J., & Gleser, G.C. (1964). The signal/noise ratio in the comparison of reliability coefficients. Educational and Psychological Measurement, 24, 467-480. (3)

Cronbach, L.J., Gleser, G.C., et al. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley. (3,13)

Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137-163. (13)

Cronbach, L.J., Schonemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. Educational and Psychological Measurement, 25, 291-312. (8)

Cronbach, L.J. (& Shavelson, R.J.) (2004). My current thoughts on coefficient alpha and successor procedures. Educational and Psychological Measurement, 64 (3), 391-418. (8)

Cudeck, R. (1980). A comparative study of indices for internal consistency. Journal of Educational Measrement, 17, 117-130. (8)

Cunny, K.A., & Perri, M. (1991). Single-item vs. multiple-item measures of health-related quality of life. Psychological Reports, 69, 127-130. (7)

Cureton, E.E. (1931). Errors of measurement and correlation. Archives of Psychology, 19, No. 125. Pp. 63. (1,2)

Cureton, E.E. (1950). Validity, reliability, and baloney. Educational and Psychological Measurement, 10, 94-96. (2)

Cureton, E.E. (1958). The definition and estimation of test reliability. Educational and Psychological Measurement, 18, 715-738. (1,3)

Cureton, E.E. (1965). Reliability and validity: Basic assumptions and experimental designs. Educational and Psychological Measurement, 25, 327-346. (2)

Darroch, J.N., & McCloud, P.I. (1986). Category distinguishability and observer agreement. Australian Journal of Statistics, 28, 371-388. (7)

Davies, M., & Fleiss, J.L. (1982). Measuring agreement for multinomial data. Biometrics, 38, 1047-1051. (7)

Davis, F.B., & Fifer, G. (1959). The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 19, 159-170. (13)

DeKeyser, F.G., & Pugh, L.C. (1990). Assessment of the reliability and validity of biochemical measures. Nursing Research, 39, 314-317. (3)

Donner, A., & Eliasziw, M. (1987). Sample size requirements for reliability studies. Statistics in Medicine, 6, 441-448. (11)

Donoghue, J.R., & Cliff, N. (1991). An investigation of ordinal true score theory. Applied Psychological Measurement, 15, 335-351. (2)

Doran, H.C. (2005). The information function for the one-parameter logistic model: Is it reliability? Educational and Psychological Measurement, 65 (5), 665-675. (13).

Dressel, P.L. (1940). Some remarks on the Kuder-Richardson reliability coefficient. Psychometrika, 5, 305-310. (8)

Drewes, D.W. (2000). Beyond the Spearman-Brown: A structural approach to maximal reliability. Psychological Methods, 5, 214-227. (13)

DuBois, D., & DuBois, E.F. (1916). A formula to estimate the approximate surface area if height and weight be known. Archives of Internal Medicine, 17, 863-871. (5)

DuBois, P.H. (1957). Multivariate correlational analysis. New York: Harper. (6)

Dudek, F.J. (1979). The continued misinterpretation of the standard error of measurement. Psychological Bulletin, 86 (2), 335-337. (5)

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. Journal of Applied Psychology, 89 (5), 792-808. (11)

Dunn, G. (2004). Statistical evaluation of measurement errors: Design and analysis of reliability studies (2nd. ed.). London: Arnold. (2,3,7,9)

Ebel, R.L. (1961). Must all tests be valid? American Psychologist, 16, 640-647. (B)

Ebel, R.L. (1965). Confidence weighting and test reliability. Journal of Educational Measurement, 2, 49-57. (13)

Ebel, R.L. (1967). The relation of item discrimination to test reliability. Journal of Educational Measurement, 4, 125-128. (8)

Ebel, R.L. (1969a). Expected reliability as a function of choices per item. Educational and Psychological Measurement, 29, 565-570. (13)

Ebel, R.L. (1969b). The relation of scale fineness to grade accuracy. Journal of Educational Measurement, 6, 217-221. (13)

Ebel, R.L. (1972). Why a longer test is usually more reliable. Educational and Psychological Measurement, 32, 249-253. (13)

Edgerton, H.A., & Toops, H.A. (1928). A formula for finding the average inter-correlation coefficient of unranked raw scores without solving any of the individual intercorrelations. Journal of Educational Psychology, 19, 131-138. (8)

Edwards, J.R. (2001). Ten difference score myths. Organizational Research Methods, 4 (3), 265-287. (6)

Ehrman, B.D. (2005). Misquoting Jesus: The story behind who changed the bible and why. New York: Harper. (E)

Eide, G.E., & Heuch, I. (2001). Attributable fractions: Fundamental concepts and their visualization. Statistical Methods in Medical Research, 10, 159-193. (F)

Eliasziw, M., & Donner, A. (1987). A cost-function approach to the design of reliability studies. Statistics in Medicine, 6, 647-655. (11)

Enders, C.K. (2003). Using the EM algorithm to estimate coefficient alpha for scales with item-level missing data. Psychological Methods, 8, 322-337. (13)

Enders, C.K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. Educational and Psychological Measurement, 64 (3), 419-436. (13)

Engstrom, F.M., Roche, A.F., & Mukherjee. (1981). Differences between arm span and stature in White children. Journal of Adolescent Health Care, 2, 19-22. (12)

Engstrom, J.L. (1988). Assessment of the reliability of physical measures. Research in Nursing & Health, 11, 383-389. (1,3, 10, D)

Engstrom, J.L., Kavanaugh, K., et al. (1995). Reliability of in-bed weighing procedures for critically ill infants. Neonatal Network, 14, 27-33. (D)

Ennis, R.H. (1999). Test reliability: A practical exemplification of ordinary language philosophy. Yearbook of the Philosophy of Education Society. (1)

Enstrom, J.E., & Kabat, G.C. (2003). Environmental tobacco smoke and tobacco related mortality in a prospective study of Californians, 1960-1998. British Medical Journal, 326, 1057-1066. (F)

Enstrom, J.E., & Kabat, G.C. (2006). Environmental tobacco smoke and coronary disease mortality in the United States: A meta-analysis and critique. Inhalation Toxicology, 18, 199-210. (F)

EPA (1992). Respiratory health effects of passive smoke: Lung cancer and other disorders. Washington, DC: Office of Health and Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency. (F)

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM Guidelines editorial. Educational and Psychological Measurement, 61, 517-531. (11)

Fauveau, V. (2006). Assessing probable causes of death without death registration or certificates: A new science? Bulletin of the World Health Organization, 84 (3), 246-247. (F)

Feinstein, A.R. (1985). Clinical epidemiology: The architecture of clinical research. Philadelphia: Saunders. (1)

Feinstein, A.R. (1987). Clinimetrics. New Haven, CT: Yale University Press. (1)

Feinstein, A.R. (1992). Critique (of Smith, Sears, et al., 1992). Toxicologic Pathology, 20 (2), 303-305. (F)

Feinstein, A.R., & Cicchetti, D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology, 43, 543-549. (7)

Feldt, L.S. (1965). The approximate sampling distribution of Kuder-Richardson [reliability] coefficient twenty. Psychometrika, 30, 357-370. (11)

Feldt, L.S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 34, 363-373. (11)

Feldt, L.S. (1975). Estimation of the reliability of a test divided into two parts of unequal length. Psychometrika, 40, 557-561. (13)

Feldt, L.S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. Psychometrika, 45, 99-105. (11)

Feldt, L.S. (1996). Confidence intervals for the proportion of mastery in criterion-referenced measurement. Journal of Educational Measurement, 33, 106-114. (13)

Feldt, L.S. (1997). Can validity rise when reliability declines? Applied Measurement in Education, 10, 377-387. (B)

Feldt, L.S. (2005). Estimating the reliability of dichotomous or trichotomous scores. Educational and Psychological Measurement, 65 (1), 28-41. (13)

Feldt, L.S., & Ankenmann, R.D. (1998). Appropriate sample size for a test of equality of alpha coefficients. Applied Psychological Measurement, 22, 170-178. (11)

Feldt, L.S., & Ankenmann, R.D. (1999). Determining sample size for a test of equality of alpha coefficients when the number of part-tests is small. Psychological Methods, 4, 366-377. (11)

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), Educational measurement (3rd ed., pp. 105-146). New York: Macmillan. (3)

Feldt, L.S., & Charter, R.A. (2003). Estimating the reliability of a test split into two parts of equal or unequal length. Psychological Methods, 8 (1), 102-109.

Feldt, L.S., & Charter, R.A. (2006). Averaging internal consistency reliability coefficients. Educational and Psychological Measurement, 66 (2), 215-227. (8)

Feldt, L.S., Woodruff, D.J., & Salih, F.A. (1987). Statistical inference for coefficient alpha. Applied Psychological Measurement, 11, 93-103. (11)

Ferketich, S. (1990). Internal consistency estimates of reliability. Research in Nursing & Health, 13, 437-440. (8)

Fisher, R.A. (1925). Statistical methods for research workers. Available for free access on the Classics in the History of Psychology (CHP) website developed by Christopher Green of York University in Canada. (9)

Fleiss, J.L. (1965). Estimating the accuracy of dichotomous judgments. Psychometrika, 30, 469-479. (7)

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 378-382. (7)

Fleiss, J.L. (1975). Measuring agreement between judges on the presence or absence of a trait. Biometrics, 31, 651-659. (7)

Fleiss, J.L. (1976). Comment on Overall and Woodward's asserted paradox concerning the measurement of change. Psychological Bulletin, 83, 774-775. (6)

Fleiss, J.L. (1986). The design and analysis of clinical experiments. New York: Wiley. (4)

Fleiss, J.L., Cohen, J., & Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 72, 323-327. (11)

Fleiss, J.L., Levin, B., & Paik, M.C. (2003). Statistical methods for rates and proportions (3rd ed.). New York: Wiley.

Fleiss, J.L., & Shrout, P.E. (1977). The effects of measurement errors on some multivariate procedures. American Journal of Public Health, 67, 1188-1191. (4)

Fleiss, J.L., & Shrout, P.E. (1978). Approximate interval estimation for a certain intraclass correlation coefficient. Psychometrika, 43, 259-262. (11)

Forbes, S., & Taunton, R.L. (1994). Reliability of aggregated organizational data: An evaluation of five empirical indices. Journal of Nursing Measurement, 2 (1), 37- 48. (13)

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18, 39-50. (13)

Franzen, R., & Derryberry, M. (1932a). Note on reliability coefficients. Journal of Educational Psychology, 23, 559-560. (10)

Franzen, R., & Derryberry, M. (1932b). Reliability of group distinctions. Journal of Educational Psychology, 23, 586-593. (10)

Fuller, W.A., & Hidiroglou, M.A. (1978). Regression estimates after correction for attenuation. Journal of the American Statistical Association, 73, 99-104. (4)

Funk, R.W., Hoover, R.W., and The Jesus Seminar. (1993). The five gospels. San Francisco: Harper. (E)

Funk, R.W., and The Jesus Seminar. (1998). The acts of Jesus. San Francisco: Harper. (E)

Furr & Bacharach (2008).

Gardner, P.L. (1970). Test length and the standard error of measurement. Journal of Educational Measurement, 7, 271-273. (8)

Gardner, R.C., & Neufeld, R.W.J. (1987). Use of the simple change score in correlational analyses. Educational and Psychological Measurement, 47, 849-862. (6)

Gift, A.G., & Soeken, K.L. (1988). Assessment of physiologic instruments. Heart & Lung, 17, 128-133. (1)

Gilmer, J.S., & Feldt, L.S. (1983). Reliability estimation for a test with parts of unknown length. Psychometrika, 48, 99-111. (13)

Gittelsohn, A.M., & Royston, P.N. (1982). Annotated bibliography of cause-of-death validation studies, 1958-1980. DHHS publication number 82-1363. Hyattsville, MD: US Dept. of Health and Human Services, Public Health Services, Office of Health Research, Statistics, and Technology, National Center for Health Statistics. (F)

Glantz, S.A., & Parmley, W.W. (1991). Passive smoking and heart disease: Epidemiology, physiology, and biochemistry. Circulation, 83, 1-12. (F)

Glantz, S.A., & Parmley, W.W. (1995). Passive smoking and heart disease: Mechanisms and risk. Journal of the American Medical Association, 273, 1047-1053. (F)

Glantz, S.A., & Parmley, W.W. (2001). Even a little secondhand smoke is dangerous. Journal of the American Medical Association, 286, 462-463. (F)

Glass, G.V (1968). Response to Traub's "Note on the reliability of residual change scores". Journal of Educational Measurement, 5, 265-267. (6)

Glass, G.V, & Wiley, D.E. (1964). Formula scoring and test reliability. Journal of Educational Measurement, 1, 43-47. (13)

Glynn, W.J., & Muirhead, R.J. (1978). Inference in canonical correlation analysis. Journal of Multivariate Analysis, 8, 468-478. (11)

Goodenough, F.L. (1936). A critical note on the use of the term "reliability" in mental measurement. Journal of Educational Psychology, 27, 173-178. (1)

Goodenough, F.L. (1949). Mental testing: Its history, principles, and applications (2nd. ed.). New York: Rinehart. (1)

Goodman, L.A., & Kruskal, W.H. (1979). Measures of association for cross classifications. New York: Springer-Verlag. (10)

Graham, J.M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. Educational and Psychological Measurement, 66 (6), 930-934. (3)

Grant, B.F., Dawson, et al. (2003). The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): Reliability of alcohol

consumption, tobacco use, family history of depression, and psychiatric diagnostic modules in a general population sample. Drug and Alcohol Dependence, 71, 7-16. (1)

Grant, B.F., Harford, T.C., et al. (1995). The alcohol use disorder and associated disabilities interview schedule (AUDADIS): Reliability of alcohol and drug modules in a general population sample. Drug and Alcohol Dependence, 39 , 37-44. (1)

Green, B.F. (1950). A note on the calculation of weights for maximum battery reliability. Psychometrika, 15, 57-61. (13)

Green, S.B. (2004). A coefficient alpha for test-retest data. Psychological Methods, 8 (1), 83-101. (8)

Green, S.B., Lissitz, R.W., & Mulaik, S.A. (1977). Limitations of coefficient alpha as an index of unidimensionality. Educational and Psychological Measurement, 37, 827-838. (8)

Greene, V.L., & Carmines, E.G. (1980). Assessing the reliability of linear composites. In K.F. Schuessler (Ed.), Sociological Methodology, 1980 (pp.160-175). San Francisco: Jossey-Bass. (8)

Greenland, P., Bowley, N.L., et al. (1990). Precision and accuracy of a portable blood analyzer system during cholesterol screening. American Journal of Public Health, 80, 181-184. (2)

Greenland, S. (2004). Bounding analysis as an inadequately specified methodology. Risk Analysis, 24 (5), 1085-1092. (F)

Gross, A.J. (1995a). Uncertainties in lung cancer risk estimates reported for exposure to environmental tobacco smoke. Environmetrics, 6 (4), 403-412. (F)

Gross, A.J. (1995b). Uncertainties in lung cancer risk estimates reported for exposure to environmental tobacco smoke revisited - A response to Bayard/Jinot/Flatman and Hanley. Environmetrics, 6 (4), 423-424. (F)

Grubbs, F.E. (1948). On estimating precision of measuring instruments and product variability. Journal of the American Statistical Association, 43, 243-264. (2)

Grubbs, F.E. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. Technometrics, 15, 53-66. (2)

Gulliksen, H. (1936). The content reliability of a test. Psychometrika, 1, 189-194. (1)

Gulliksen, H. (1945). The relation of item difficulty and interitem correlation to test variance and reliability. Psychometrika, 10, 79-91. (8)

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley. (1,2,3,5,6,12)

Gulliksen, H. (1953). Comments on Guttman's review of Theory of mental tests. Psychometrika, 18, 131-133. (3)

Gustafson, P. (2004). Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments. Boca Raton, FL: Chapman & Hall/CRC. (2)

Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika, 10, 255-282. (3,8)

Guttman, L. (1946). The test-retest reliability of qualitative data. Psychometrika, 11, 81-95. (7, 10)

Guttman, L. (1953a). A special review of Harold Gulliksen's Theory of mental tests. Psychometrika, 18, 123-130. (3)

Guttman, L. (1953b). Reliability formulas that do not assume experimental

Guyatt, G., Kirshner, B., & Jaeschke, R. (1992). Measuring health status: What are the necessary measurement properties? Journal of Clinical Epidemiology, 45 (12), 1347-1351. (6)

Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. Journal of Chronic Diseases, 47 (2), 171-178. (6)

Haber, M., & Banrhart, H.X. (2008). A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurements. Statistical Methods in Medical Research, 17, 151-170. (3)

Habermas, G.R. (2005). Recent perspectives on the reliability of the gospels. Christian Research Journal, 28 (1). (E)

Ha-Duong, M., Casman, E.A., & Morgan, M.G. (2004). Bounding poorly characterized risks: A lung cancer example. Risk Analysis, 24 (5), 1071-1083. (F)

Haertel, E.H. (2006). Reliability. In R.L. Brennan (Ed.), Educational measurement (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger. (1, 3)

Hakstian, A.R., Schroeder, M.L., & Rogers, W.T. (1988). Inferential procedures for correlation coefficients corrected for attenuation. Psychometrika, 53, 27-43. (11)

Hakstian, A.R., & Whalen, T.E. (1976). A k-sample significance test for independent alpha coefficients. Psychometrika, 41, 219-231. (11)

Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice, 12 (3), 38-45. (13)

Hambleton, R.K., & Slater, S. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting practices. Applied Measurement in Education, 10, 19-38. (13)

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage. (3,13)

Hanamura, R.C. (1975). Measuring imprecisions of measuring instruments. Technometrics, 17, 299-302. (2)

Hancock, G.R., & Mueller, R.O. (2000). Rethinking construct reliability within latent variable systems. Chapter 10 (pp. 195-216) in R. Cudek, S. duToit, and D. Sorbom (Eds.), Structural equation modeling: Present and future (A Festschrift in honor of Karl Joreskog). (13)

Hanley, J.A. (1995). Uncertainties in lung cancer risk estimates reported for exposure to environmental tobacco smoke - A rejoinder to Gross. Environmetrics, 6 (4), 419-421. (F)

Hanley, J.A. (2001). A heuristic approach to the formulas for population attributable fraction. Journal of Epidemiology and Community Health, 55, 508-514. (F)

Harris, C.W. (Ed.) (1963). Problems in measuring change. Madison, WI: University of Wisconsin Press. (6)

Harris, C.W. (1973). Note on the variances and covariances of three error types. Journal of Educational Measurement, 10, 49-50. (5,13)

Harris, J.A. (1913). On the calculation of intraclass and interclass correlations from class moments when the number of possible combinations is large. Biometrika, 9, 446-472. (9)

Hasin, D., Carpenter, K.M., et al. (1997). The alcohol use disorder and associated disabilities interview schedule (AUDADIS): Reliability of alcohol and

drug modules in a clinical sample. Drug and Alcohol Dependence, 44, 133-141. (1)

Heise, D.R. (1969). Separating reliability and stability in test-retest correlation. American Sociological Review, 34, 93-101. (6)

Heise, D.R., & Bohrnstedt, G.W. (1970). Validity, invalidity, and reliability. In E. Borgatta and G.W. Bohrnstedt (Eds.), Sociological methodology 1970 (pp. 104-129). San Francisco: Jossey-Bass. (2,8)

Heller, R.F., Buchan, I., et al. (2003). Communicating risks at the population level: Application of population impact numbers. British Medical Journal, 327, 1162-1165. (F)

Hernan, M.A. (2004). A definition of causal effect for epidemiological research. Journal of Epidemiology and Community Health, 58, 265-271. (F)

Hernan, M.A., & Robins, J.M. (2006). Estimating causal effects from epidemiological data. Journal of Epidemiology and Community Health, 60, 578-586. (F)

Hershberger, S.L., Fisher, D.G., et al. (2005). Effect of correlated errors on the test-retest reliability of self-reported age of first drug use. Paper presented at the annual meeting of the American Public Health Association, Philadelphia, PA. (3)

Hertzog, C., & Nesselroade, J.R. (2003). Assessing psychological change in adulthood: An overview of methodological issues. Psychology and Aging, 18 (4), 639-657. (6)

Himes, J.H. (1989). Reliability of anthropometric methods and replicate measurements. American Journal of Physical Anthropology, 79, 77-80. (D)

Hoffman, P.J. (1963). Test reliability and practice effects. Psychometrika, 28 (3), 273-288. (3)

Holland, P.W. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81 (396), 945-970. [Includes comments by D.B. Rubin, D.R. Cox, C.Glymour, and C.Granger, and a rejoinder by Holland.] (F)

Holland, P.W. (1993). Which comes first, cause or effect? In G. Keren & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: Methodological issues, pp. 273-282. Mahwah, NJ: Erlbaum. (F)

Holland, P.W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a nonparallel test. Psychometrika, 68 (1), 123-149. (13)

Holzinger, K.J. (1932). The reliability of a single test item. Journal of Educational Psychology, 23, 411-417. (7)

Horst, P. (1954). The estimation of immediate retest reliability. Educational and Psychological Measurement, 14, 705-708. (3)

Hoyt, C. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160. (8)

Hubert, L.J. (1977). Kappa revisited. Psychological Bulletin, 84, 289-297. (7)

Huffman, G.B. (October 1,1997). Death certificates: Why it matters how your patient died. American Family Physician. (F)

Hummel-Rossi, B., & Weinberg, S.L. (1975). Practical guidelines in applying current theories to the measurement of change. Journal Supplement Abstract Service-Catalog of Selected Documents in Psychology, 5, 226 (MS No. 916). (6)

Humphreys, L.G. (1956). The normal curve and the attenuation paradox in test theory. Psychological Bulletin, 53, 472-476. (B)

Hutchinson, T.P. (1993). Kappa muddles together two sources of disagreement: Tetrachoric correlation is preferable. Research in Nursing & Health, 16, 313-315. (7)

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 13, 253-264. (13)

Huynh, H. (1986a). On the reliability of an extreme score. Psychometrika, 51, 475-478. (5)

Huynh, H. (1986b). Estimation of the KR20 reliability coefficient when the data are incomplete. British Journal of Mathematical and Statistical Psychology, 39, 69-78. (13)

Huynh, H., & Saunders, J.C. (1980). Accuracy of two procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 17, 351-358. (13)

Jackson, R.W.B. (1939). The reliability of mental tests. British Journal of Psychology, 29, 267-287. (3)

Jackson, R.W.B. (1942). Note on the relationship between internal consistency and test-retest estimates of the reliability of a test. Psychometrika, 7, 157-164.

Jacobsen, P.B., Donovan, K.A., et al. (2005). Screening for psychologic distress in ambulatory cancer patients: A multicenter evaluation of the Distress Thermometer. Cancer, 103, 1494-1502. (7)

Jaech, J.L. (1985). Statistical analysis of measurement errors. New York: Wiley. (2)

Jarjoura, D. (1985). Tolerance intervals for true scores. Journal of Educational Statistics, 10, 1-17. (5)

Joe, G.W., & Mendoza, J.L. (1989). The internal correlation: Its applications in statistics and psychometrics. Journal of Educational Statistics, 14, 211-226. (8)

Johnson, H.G. (1944). An empirical study of the influence of errors of measurement upon correlation. American Journal of Psychology, 57, 521-536. (4)

Johnson, H.G. (1950). Test reliability and correction for attenuation. Psychometrika, 15, 115-119. (4)

Johnson, J.M. (2005). Visual analog scales: Part 1 and Part 2. Retrieved on April 12, 2005 from the following website: http://www.DCIResearchReview.htm. (7)

Johnson, R.L., Penny, J., et al. (2003). Score resolution: An investigation of the reliability and validity of resolved scores. Applied Measurement in Education, 16 (4), 299-322.

Johnson, T.S., & Engstrom, J.L. (2002). State of the science in measurement of infant size at birth. Newborn and Infant Nursing Reviews, 2, 150-158. (D)

Johnson, T.S., Engstrom, J.L., & Gelhar, D.K. (1997). Intra- and inter-examiner reliability of anthropometric measurements of term infants. Journal of Pediatric Gastroenterology and Nutrition, 24 (5), 497-505. (D)

Johnson, T.S., Engstrom, J.L., et al. (1998). Reliability of length measurements in term infants. Journal of Obstetric, Gynecologic, and Neonatal Nursing, 27 (3), 270-276. (D)

Johnson, T.S., Engstrom, J.L., et al. (1999). Reliability of three length measurement techniques in term infants. Pediatric Nursing, 25 (1), 13-17. (D)

Johnson, T.P., & Mott, J.A. (2001). The reliability of self-reported age at onset of tobacco, alcohol and illicit drug use. Addiction, 96, 1187-1198.

Jones, P.R.M., & Rioux, M. (1997). Three-dimensional surface anthropometry: Applications to the human body. Optics and Laser Engineering, 28, 89-117. (D)

Joreskog, K.G., & Sorbom, D. (1979). Advances in factor analysis and structural equation models. Cambridge, MA: Abt Books. (13)

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141-151. (8)

Kaiser, H.F. (1991). Coefficient alpha for a principal component and the Kaiser-Guttman rule. Psychological Reports, 68, 853-858. (8)

Kaiser, H.F., & Carter, H.D. (1971). A geometric representation of the notions of reliability, relevance, and validity. California Journal of Educational Research, 22 (3), 122-124. (2)

Kane, M., & Case, S.M. (2004). The reliability and validity of weighted composite scores. Applied Measurement in Education, 17 (3), 221-240. (13)

Karger, B., de la Grandmaison, L., et al. (2004). Analysis of 155 consecutive forensic exhumations with emphasis on undetected homicides. International Journal of Legal Medicine, 118, 90- 94. (F)

Kavanaugh, K.L., Meier, P.P., & Engstrom, J.L. (1989). Reliability of weighing procedures for preterm infants. Nursing Research, 38, 178-179. (D)

Kavanaugh, K.L., Engstrom, J.L., et al. (1990). How reliable are scales for weighing preterm infants? Neonatal Network, 9 (3), 29-32. (D)

Kelley, T.L. (1921). The reliability of test scores. Journal of Educational Research, 3, 370-379. (1,3)

Kelley, T.L. (1923). Statistical method. New York: Macmillan. (3,4)

Kelley, T.L. (1925). The applicability of the Spearman-Brown formula for the measurement of reliability. Journal of Educational Psychology, 16, 300-303. (3)

Kelley, T.L. (1927). The interpretation of educational measurements. New York: World Book. (3,5)

Kelley, T.L. (1942). The reliability coefficient. Psychometrika, 7, 75-83. (1,3,4)

Kelley, T.L. (1947). Fundamentals of statistics. Cambridge, MA: Harvard University Press. (3)

Kendall, M.G., & Gibbons, J.D. (1990). Rank correlation methods (5th ed.). New York: Oxford University Press. (10)

Kerlinger, F.N. (1976). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart, and Winston. (8)

King, G., & Lu, Y. (2006). Verbal autopsy methods with multiple causes of death. Retrievable at http://GKing.Harvard.edu/files/abs/vamc-abs.shtml. (F)

Klein, D.F., & Cleary, T.A. (1967). Platonic true scores and error in psychiatric rating scales. Psychological Bulletin, 68, 77-80. (7)

Klein, D.F., & Cleary, T.A. (1969). Platonic true scores: Further comment. Psychological Bulletin, 71, 278-280. (7)

Klepeis, N.E., Ott, W.R., & Switzer, P. (2007). Real-time measurement of outdoor tobacco smoke particles. Journal of the Air and Waste Management Association, 57, 522-534. (F)

Knapp, T.R. (1970). N vs. N-1. American Educational Research Journal, 7, 625-626. (3)

Knapp, T.R. (1971). Statistics for educational measurement. Scranton, PA: Intext. (3,5)

Knapp, T.R. (1977a). The unit-of-analysis problem in applications of simple correlation analysis to educational research. Journal of Educational Statistics, 2, 171-196. (13)

Knapp, T.R. (1977b). The reliability of a dichotomous test item: A correlationless approach. Journal of Educational Measurement, 14, 237-252. (7)

Knapp, T.R. (1980). The (un)reliability of change scores in counseling research. Measurement and Evaluation in Guidance, 13, 149-157. (6)

Knapp, T.R. (1984a). A respose to Williams and Zimmerman. Measurement and Evaluation in Guidance, 17, 183-184. (6)

Knapp, T.R. (1984b). The unit of analysis and the independence of observations. Undergraduate Mathematics and its Applications Project (UMAP) Journal, 5, 363-388. (13)

Knapp, T.R. (1985). Validity, reliability, and neither. Nursing Research, 34, 189-192. (1,2,B)

Knapp, T.R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. Nursing Research, 39, 121-123. (10)

Knapp, T.R. (1991). Coefficient alpha: Conceptualizations and anomalies. Research in Nursing & Health, 14, 457-460. See also Errata, op. cit., 1992, 15, 321. (8,11)

Knapp, T.R. (1992). Technical error of measurement: A methodological critique. American Journal of Physical Anthropology, 87, 235-236. (3)

Knapp, T.R. (1993). Treating ordinal scales as ordinal scales. Nursing Research, 42, 184-186. (10)

Knapp, T.R. (1998). Quantitative nursing research. Thousand Oaks, CA: Sage. [Now accessible free of charge at www.tomswebpage.net.] (1)

Knapp, T.R. (1999). The analysis of the data for two-way contingency tables. Research in Nursing & Health, 22, 263-268.

Knapp, T.R. (2001). Reporting the reliability of research instruments. Nurse Author & Editor, 11 (3), 1-2, 4. (3)

Knapp, T.R. (2007). Sex, age, height, and weight. Accessible free of charge at www.tomswebpage.net. (D)

Knapp, T.R., & Brown, J.K. (1995). Ten measurement commandments that often should be broken. Research in Nursing & Health, 18, 465-469. (1,7,8)

Knapp, T.R., Kimble, L.P., & Dunbar, S.B. (1998). Distinguishing between the stability of a construct and the stability of an instrument in trait/state measurement. Nursing Research, 47, 60-62. (6)

Knapp, T.R., & Sawilowsky, S.S. (2001a). Constructive criticisms of methodological and editorial practices. Journal of Experimental Education. (1)

Knapp, T.R., & Sawilowsky, S.S. (2001b). Strong arguments: Rejoinder to Thompson. Journal of Experimental Education. (1)

Knapp, T.R., & Tam, H.P. (2007). Is true score a latent variable? Accessible free of charge at www.tomswebpage.net. (2)

Koning, A.J., & Franses, P.H. (2003). Confidence intervals for Cronbach's coefficient alpha values. ERIM Report Series ERS-2003-041-MKT. (11)

Kraemer, H.C. (1981). Extension of Feldt's approach to testing homogeneity of coefficients of reliability. Psychometrika, 46 (1), 41-45. (11)

Kristof, W. (1963a). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. Psychometrika, 28, 221-238. (11)

Kristof, W. (1963b). Statistical inferences about the error variance. Psychometrika, 28, 129-143. (11)

Kristof, W. (1964). Testing differences between reliability coefficients. British Journal of Statistical Psychology, 17, 105-111. (11)

Kristof, W. (1970). On sampling theory of reliability estimation. Journal of Mathematical Psychology, 7, 371-377. (11)

Kristof, W. (1971). On the theory of a set of tests which differ only in length. Psychometrika, 36, 207-225. (13)

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. Psychometrika, 39, 491-499. (13)

Krus, D.J., & Helmstadter, G.C. (1987). The relationship between correlational and internal consistency notions of test reliability. Educational and Psychological Measurement, 47, 911-915. (8)

Krus, D.J., & Helmstadter, G.C. (1993). The problem of negative reliabilities. Educational and Psychological Measurement, 53, 643-650. (8)

Kuder, G.F. (1991). Comments concerning the appropriate use of formulas for estimating the internal-consistency reliability of tests. Educational and Psychological Measurement, 51, 873-874. (8)

Kuder, G.F., & Richardson, M.W. (1937). The theory of the estimation of test reliability. Psychometrika, 2, 151-160. (8)

Kuehn, B. (2006). Report reviews secondhand smoke risks: Some scientists question risk level. Journal of the American Medical Association, 296 (8), 922-923. (F)

Kwok, T., & Whitelaw, M.N. (1991). The use of arm span in nutritional assessment of the elderly. Journal of the American Geriatrics Society, 39, 492-496. (12)

Labouvie, E., Bates, M.E., & Pandina, R.J. (1997). Age of first use: Its reliability and predictive utility. Journal of Studies on Alcohol & Drugs, 58, 638-643. (10)

LaForge, R. (1965). Components of reliability. Psychometrika, 30, 187-195. (8)

Landis, J.R., & Koch, G.C. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174. (7)

Larson, M.R. (2000). Social desirability and self-reported weight and height. International Journal of Obesity and Related Metabolic Disorders, 24, 663-665. (12)

Laschinger, H.K.S. (1992). Intraclass correlations as estimates of interrater reliability in nursing research. Western Journal of Nursing Research, 14, 246-251. (9)

Last, J.M. (2001). A dictionary of epidemiology (4th ed.). New York: Oxford University Press. (1)

Laveault, D., Zumbo, B.D., et al. (Eds.) (1994). Modern theories of measurement: Problems and issues. Ottawa: University of Ottawa. (6)

LeBreton, J. M., Burgess, J. R., et al. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? Organizational Research Methods, 6 (1), 80-128. (8)

Lee, P.S.C., & Suen, H.K. (1984). The estimation of kappa from percentage agreement interobserver reliability. Behavioral Assessment, 6, 375-378. (7)

Lee, R., Miller, K., & Graham, W. (1982). Corrections for restriction of range and attenuation in criterion-related validation studies. Journal of Applied Psychology, 67, 637-639. (4)

Lenfant, C., Friedman, L., & Thom, T. (1998). Fifty years of death certificates: The Framingham Heart Study. Annals of Internal Medicine, 129, 1066-1067. (F)

Levin, M.L. (1953). The occurrence of lung cancer in man. Acto Unio Internationalis Contra Cancrum, 9, 531-541. (F)

Li, H. (1997). A unifying expression for the maximal reliability of a linear composite. Psychometrika, 62, 245-249. (13)

Li, H. (2003). The resolution of some paradoxes related to reliability and validity. Journal of Educational and Behavioral Statistics, 28 (2), 89-95.

Li, H., Rosenthal, R., & Rubin, D.B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. Psychological Methods, 1, 98-107. (13)

Li, H., & Wainer, H. (1997). Toward a coherent view of reliability in test theory. Journal of Educational and Behavioral Statistics, 22, 478-484. (8)

Liao, J.J.Z. (2003). An improved concordance correlation coefficient. Pharmaceutical Statistics, 2, 253-261. (2)
Lin, L.I. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45, 255-268. (2)

Lincoln, E.A. (1932). The unreliability of reliability coefficients. Journal of Educational Psychology, 23, 11-14. (1,10)

Lincoln, E.A. (1933). Reliability coefficients are still unreliable. Journal of Educational Psychology, 24, 235-236. (1,10)

Linn, R.L. (1994). Performance assessment: Policy promise and technical measurement standards. Educational Researcher, 23 (9), 4-14. (2)

Linn, R.L., & Slinde, J.A. (1977). The determination of the significance of change between pre- and posttesting periods. Review of Educational Research, 47, 121-150. (6)

Liou, M. (1989). A note on reliability estimation for a test with components of unknown functional lengths. Psychometrika, 54, 153-163. (13)

Little, R.J.A., & Rubin, D.B. (2002). Statistical analysis with missing data (2nd ed.). New York: Wiley. (13)

Livingston, S.A. (1972). Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 9, 13-26. (13)

Livingston, S.A. (1973). A note on the interpretation of the criterion-referenced reliability coefficient. Journal of Educational Measurement, 10, 311. (13)

Livingston, S.A. (2004). An interesting problem in the estimation of scoring reliability. Journal of Educational and Behavioral Statistics, 29 (3), 333-341. (1)

Livingston, S.A., & Wingersky, M.S. (1979). Assessing the reliability of tests used to make pass/fail decisions. Journal of Educational Measurement, 16, 247-260. (13)

Lloyd-Jones, D.M., Martin, D.O., et al. (1998). Accuracy of death certificates for coding coronary heart disease as the cause of death. Annals of Internal Medicine, 129, 1020-1026. (F)

Loevinger, J. (1954). The attenuation paradox in test theory. Psychological Bulletin, 51, 493-504. (B)

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635-694. (2,3)

Lord, F.M. (1944). Reliability of multiple-choice tests as a function of number of choices per item. Journal of Educational Psychology, 35, 175-180. (13)

Lord, F.M. (1952). A theory of test scores. Psychometric Monograph No. 7. (13)

Lord, F.M. (1955). Estimating test reliability. Educational and Psychological Measurement, 15, 325-336. (8)

Lord, F.M. (1956). The measurement of growth. Educational and Psychological Measurement, 16, 421-437. (6)

Lord, F.M. (1957a). Do tests of the same length have the same standard error of measurement? Educational and Psychological Measurement, 17, 510-521. (8)

Lord, F.M. (1957b). A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. Psychometrika, 22, 207-220. (11)

Lord, F.M. (1958). Further problems in the measurement of growth. Educational and Psychological Measurement, 18, 437-451. (6)

Lord, F.M. (1959a). Tests of the same length do have the same standard error of measurement. Educational and Psychological Measurement, 19, 233-239. (8)

Lord, F.M. (1959b). Statistical inferences about true scores. Psychometrika, 24, 1-17. (5)

Lord, F.M. (1959c). An approach to mental test theory. Psychometrika, 24, 283-302. (2,3)

Lord, F.M. (1959d). Inferences about true scores from parallel test forms. Educational and Psychological Measurement, 19, 331-336. (5)

Lord, F.M. (1964). Nominally and rigorously parallel test forms. Psychometrika, 29, 335-346. (12)

Lord, F.M. (1973). Testing if two measuring procedures measure the same dimension. Psychological Bulletin, 79, 71-72.

Lord, F.M. (1974). Variance stabilizing transformation of the stepped-up reliability coefficient. Journal of Educational Measurement, 11, 55-57.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum. (13)
Lord, F.M. (1984). Standard errors of measurement at different ability levels. Journal of Educational Measurement, 21, 239-243. (5)

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley. (2,7)

Lumsden, J. (1976). Test theory. Annual Review of Psychology, 27, 251-280. (2,3)

Lynn, M.R. (1989). Instrument reliability and validity: How much needs to be published? Heart & Lung, 18, 421-423. (1)

Maassen, G.H. (2000). Kelley's formula as a basis for the assessment of reliable change. Psychometrika, 65 (2), 187-197. (6)

Maberly, N.C. (1967). Characteristics of internal consistency estimates within restricted score ranges. Journal of Educational Measurement, 4, 15-28. (8)

MacCallum, R.C., Zhang, S., et al. (2002). On the practice of dichotomization of quantitative variables. Psychological Methods, 7 (1), 19-40. (13)

Maclure, M., & Willett, W.C. (1987). Misinterpretation and misuse of the Kappa statistic. American Journal of Epidemiology, 126, 161-169. (7)

Magrane, B.P., Gilliland, M.G.F., & King, D.E. (1997). Certification of death by family physicians. American Family Physician, 56, 1433-1438. (F)

Malgady, R.G., & Colon-Malgady, G. (1991). Comparing the reliability of difference scores and residuals in analysis of covariance. Educational and Psychological Measurement, 51, 803-807. (6)

Manning, W.H., & DuBois, P.H. (1962). Correlational methods in research on human learning. Perceptual and Motor Skills, 15, 287-321. (6)

Marascuilo, L.A., & McSweeney, M. (1977). Nonparametric and distribution-free methods for the social sciences. Monterey, CA: Brooks/Cole. (10)

Marcus-Roberts, H.M., & Roberts, F.S. (1987). Meaningless statistics. Journal of Educational Statistics, 12, 383-394. (10)

Maris, E. (1998). Covariance adjustment versus gain scores---revisited. Psychological Methods, 3, 309-327. (6)

Marks, G.C., Habicht, J.P., & Mueller, W.H. (1989). Reliability, dependability, and precision of anthropometric measurements: The Second National Health and Nutrition Examination Survey, 1976-1980. American Journal of Epidemiology, 130, 578-587. (1, D)

Marradi, A. (1990). Reliability: A dissenting view. Bulletin Methodologie Sociologique, 28, 56-71. (1)

Mathers, C.D., Fat, D.M., et al. (2005). Counting the dead and what they died from: An assessment of the global status of cause of death data. Bulletin of the World Health Organization, 83, 171-177. (F)

Mattson, D. (1965). The effects of guessing on the standard error of measurement and the reliability of test scores. Educational and Psychological Measurement, 25, 727-730. (13)

Maxwell, A.E. (1961). Analysing qualitative data. London: Methuen. (10)

Maxwell, A.E. (1968). The effect of correlated errors on estimates of reliability coefficients. Educational and Psychological Measurement, 28, 803-811. (3)

Maxwell, S. E., & Howard, G.S. (1981). Change scores--Necessarily anathema? Educational and Psychological Measurement, 41, 747-756. (6)

McCall, W.A. (1923). How to measure in education. New York: Macmillan. (1,2)

McDonald, R.P. (1999). Test theory: A unified treatment. Mahwah, NJ: Erlbaum. (3,8,13)

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1, 30-46. (9,11)

McNemar, Q. (1958). On growth measurement. Educational and Psychological Measurement, 18, 47-55. (6)

Meier, P.P., Engstrom, J.L., et al. (1994). A new scale for in-home test weighing for mothers of preterm and high risk infants. Journal of Human Lactation, 10, 163-166. (D)

Meier, P.P., Lysakowski, T.Y., et al. (1990). The accuracy of test weighing for preterm infants. Journal of Pediatric Gastroenterology and Nutrition, 10, 62-65. (D)

Mendoza, J.L., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. Journal of Educational Statistics, 12, 282-293. (4)

Mendoza, J.L., Stafford, K.L., & Stauffer, J.M. (2000). Large-sample confidence intervals for validity and reliability coefficients. Psychological Methods, 5, 356-369. (11)

Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). Washington, DC: American Council on Education. (B)

Messite, J., & Stellman, S.D. (1996). Accuracy of death certificate completion: The need for formalized physician training. Journal of the American Medical Association, 275 (10), 794-796. (F)

Miettinen, O.S. (1974). Proportion of disease caused or prevented by a given exposure, trait or intervention. American Journal of Epidemiology, 99, 325-332. (F)

Miller, M.B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. Structural Equation Modeling, 2, 255-273. (13)

Mislevy, R.J. (1996). Test theory reconceived. Journal of Educational Measurement, 33 (4), 379-416. (13)

Mislevy, R.J. (2004). Can there be reliability without "reliability"? Journal of Educational and Behavioral Statistics, 29 (2), 241-244. (2)

MMWR (December 14, 2001). State-specific prevalence of current cigarette smoking among adults, and policies and attitudes about secondhand smoke--- United States, 2000. Morbidity and Mortality Weekly Report. (F)

Mollenkopf, W.G. (1949). Variation of the standard error of measurement. Psychometrika, 14, 189-229. (5)

Mosier, C.I. (1943). On the reliability of a weighted composite. Psychometrika, 8, 161-168. (6,13)

Moss, P.A. (1994). Can there be validity without reliability? Educational Researcher, 23 (2), 5-12. (2)

Moss, P.A. (2004). The meaning and consequences of "reliability". Journal of Educational and Behavioral Statistics, 29 (2), 245-249. (2)

Muchinsky, P.M. (1996). The correction for attenuation. Educational and Psychological Measurement, 56, 63-75. (4)

Mueller, R.O. (1996). Basic principles of structural equation modeling: An introduction to LISREL and EQS. New York: Springer. (3,13)
Murdaugh, C. (1981). Measurement error and attenuation. Western Journal of Nursing Research, 3, 252-256. (4)

Murphy, K.R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. Personnel Psychology, 53 (4), 873-900. (8)

Nachman, G. (April 19, 1991). Where there's smoking there's ire. San Francisco Chronicle. (F)

Nashelsky, M. B., & Lawrence, C. H. (2003). Accuracy of cause of death determination without forensic autopsy examination. American Journal of Forensic Medicine and Pathology, 24 (4), 313-319. (F)

Nicewander, W.A., & Price, J.M. (1978). Dependent variability reliability and the power of significance tests. Psychological Bulletin, 85, 405-409. (11)

Nilsson, R. (2001). Environmental tobacco smoke revisited: The reliability of the data used for risk assessment. Risk Analysis, 21 (4), 737-760. (F)

Norris, S.N. (1999). Trustworthiness and consistency. Yearbook of the Philosophy of Education Society. (1)

Northam, S. & Knapp, T.R. (2006). The reliability and validity of birth certificates. Journal of Obstetric, Gynecologic, and Neonatal Nursing, 35, 3-12. (C)

Northam, S., & Knapp, T.R. The reliability and validity of death certificates. Unpublished paper, University of Texas, Tyler. (C)

Novick, M.R. (1966). The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3, 1-18. (3)

Novick, M.R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. Psychometrika, 32, 1-13. (8)

Nunnally, J.C., & Bernstein, I.H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill. (8)

O'Connor, E.F. (1972). Extending classical test theory to the measurement of change. Review of Educational Research, 42, 73-97. (6)

Okoli, C. T.C., Hall, L.A., et al. (2007). Measuring tobacco smoke exposure among smoking and nonsmoking bar and restaurant workers. Biological Research in Nursing, 9, 81-90. (F)

Osburn, H.G. (2000). Coefficient alpha and related internal consistency reliability coefficients. Psychological Methods, 5, 343-355. (8)

OSH/CDC (October, 2006). Fact sheet: Secondhand smoke causes lung cancer. Office of Smoking and Health, Centers for Disease Control and Prevention. Retrievable at www.cdc.gov/tobacco/data_statistics/Factsheets. (F)

Oumlil, A.B., & Balloun, J.L. (1986). Improving the accuracy of measurement in attitudinal and demographic models: The application of the correction for attenuation method. Journal of Business Research, 14, 355-369. (4)

Overall, J.E., & Woodward, J.A. (1975). Unreliability of difference scores: A paradox for measurement of change. Psychological Bulletin, 82, 85-86. (6,11)

Overall, J.E., & Woodward, J.A. (1976). Reassertion of the paradoxical power of tests of significance based on unreliable difference scores. Psychological Bulletin, 83, 776-777. (11)

Parascandola, M. (1998). What is wrong with the probability of causation? Jurimetrics Journal, 39, 29-44. (F)

Parker, J.M., Dillard, T.A., & Phillips, Y.Y. (1996). Arm span-height relationships in patients referred for spirometry. American Journal of Respiratory and Critical Care Medicine, 154, 533-536. (12)

Payne, R.W. (1989). Reliability theory and clinical psychology. Journal of Clinical Psychology, 45, 351-352. (5)

Payne, W.H., & Anderson, D.E. (1968). Significance levels for the Kuder-Richardson Twenty: An automated sampling experiment approach. Educational and Psychological Measurement, 28, 23-39. (11)

Pearl, J. (2000). Causality. New York: Cambridge University Press. (F)

Pearson, K. (1904). On the laws of inheritance in man. II. On the inheritance of the mental and moral characters in man, and its comparison with the inheritance of physical characters. Biometrika, 3, 131-190. (9)

Peng, C.J., & Subkoviak, M.J. (1980). A note on Huynh's normal approximation procedure for estimating criterion referenced reliability. Journal of Educational Measurement, 17, 359-368. (13)

Pirie, P., Jacobs, D., et al. (1981). Distortion in self-reported height and weight data. Journal of the American Dietetic Association, 78, 601-606. (12)

Pirkle, J.L., Bernert, J.L., et al. (2006). Trends in the exposure of nonsmokers in the U.S. population to secondhand smoke, 1988-2002. Environmental Health Perspectives, 114 96), 153-158. (F)

Plumlee, L.B. (1952). The effect of difficulty and chance success on item-test correlations and test reliability. Psychometrika, 17, 69-86. (13)

Plumlee, L.B. (1954). The predicted and observed effect of chance success on multiple-choice test validity. Psychometrika, 19, 65-70. (13)

Pothoff, E.F., & Barnett, N.E. (1932). Comparison of marks based upon weighted and unweighted items in new-type examinations. Journal of Educational Psychology, 23, 92-98. (13)

Puhan, G., & Gall, L. (2005). Reliability of pass/fail decisions on a large-scale certification test. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada. (13)

Rae, G. (2006). Correcting coefficient alpha for correlated errors: Is α_k a lower bound to reliability? Applied Psychological Measurement, 30 (1), 56-59. (8)

Raju, N.S. (1977). A generalization of coefficient alpha. Psychometrika, 42, 549-565. (8)

Raju, N.S., Lezotte, D.V., & Fearing, B.K. (2006). A note on correlations corrected for unreliability and range restriction. Applied Psychological Measurement, 30 (2), 145-149. (4)

Raju, N.S., Price, L.R., & Oshima, T.C. (2005). Conditional reliability. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada. (5)

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research. (13)

Raudenbush, S.W., & Bryk, A.S. (2002). Hierarchical linear models: Applications and data analysis methods (2^{nd} ed.). Newbury Park, CA: Sage. (13)

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. Applied Psychological Measurement, 21, 173-184. (13)

Ree, M.J., & Carretta, T.R. (2006). The role of measurement error in familiar statistics. Organizational Research Methods, 9 (1), 99-112. (4)

Renzo, J.C.A. (2002). The agreement between two diagnostic methods in binary cases: a proposal. Scandinavian Journal of Clinical Laboratory Investigations, 62, 391-398. (7)

Renzo, J.C.A. (2003). Failures of common measures of agreement in medicine and the need for a better tool: Feinstein's paradoxes and the dual vision method. Scandinavian Journal of Clinical Laboratory Investigations, 63, 207-216. (7)

Repace, J.L., & Lowrey, A.H. (1985). A quantitative estimate of nonsmokers' lung cancer risk from passive smoking. Environment International, 11, 3-22. (F)

Rippey, R. (1968). Probabilistic testing. Journal of Educational Measurement, 5, 211-215. (13)

Rippey, R. (1970). A comparison of five different scoring functions for confidence tests. Journal of Educational Measurement, 7, 165-170. (13)

Roberts, M.D. (2007). Can we trust the gospels?: Investigating the reliability of Matthew, Mark, Luke, and John. Wheaton, IL: Crossway Books. (E)

Swaen, G., & Amelsvoort. (2009). A weight of evidence approach to causal inference. Journal of Clinical Epidemiology, 62, 270-277.

Robins, J. (May 4, 2004). Should compensation schemes be based on the probability of causation or expected years of life? Web essay. (F)

Robinson, W.S. (1957). The statistical measurement of agreement. American Sociological Review, 22, 17-25.

Rockette, H.E. (1993). What evidence is needed to link lung cancer and second-hand smoke? Chance, 6 (4), 15-18. (F)

Rodacki, C.L., Fowler, N.E., et al. (2001). Technical note: Repeatability of measurement in determining stature in sitting and standing postures. Ergonomics, 44 (12), 1076-1085. (D)

Rogers, W.M., Schmitt, N., & Mullins, M.E. (2002). Correction for unreliability of multifactor measures: Comparison of alpha and parallel forms approaches. Organizational Research Methods, 5 (2), 184-199. (4,8)

Rogosa, D.R. (1988). Myths about longitudinal research. In K.W. Schaie, R.T. Campbell, W.M. Meredith, and S.C. Rawlings (Eds.), Methodological issues in aging research (pp. 171-209). New York: Springer. Reprinted in the following reference. (6)

Rogosa, D.R. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J.M. Gottman (Ed.), The analysis of change (pp. 3-66). Mahwah, NJ: Erlbaum. (6)

Rogosa, D.R. (2002) Shoe shopping and the reliability coefficient. Educational Assessment, 7 (4), 254-257. (3, D)

Rogosa, D.R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. Psychological Bulletin, 90, 726-748. (6)

Rogosa, D.R., & Willett, J.B. (1983). Demonstrating the reliability of the difference score in the measurement of change. Journal of Educational Measurement, 20, 335-343. (6)

Rosenberg, S.N., Verzo, B., et al. (1992). Reliability of length measurements for preterm infants. Neonatal Network, 11 (2), 23-27. (D)

Rosenthal, R., & Rosnow, R.L. (1991). Essentials of behavioral research: Methods and data analysis. New York: McGraw-Hill. (3)

Rosner, B., & Gore, R. (2001). Measurement error correction in nutritional epidemiology based on individual foods, with application to the relation of diet to breast cancer. American Journal of Epidemiology, 154, 827-835. (2)

Ross, J., & Lumsden, J. (1968). Attribute and reliability. British Journal of Mathematical and Statistical Psychology, 21, Part 2, 251-263. (2,3)

Roth, A.J., Kornblith, A.B., et al. (1998). Rapid screening for psychologic distress in men with prostate carcinoma. Cancer, 82, 1904-1908. (7)

Rothman, K.J., & Greenland, S. (1998). Modern epidemiology. (2nd ed.). Philadelphia: Lippincott, Williams, & Wilkins. (F)

Rowe, A.K., Powell, K.E., & Flanders, W.D. (2004). Why population attributable fractions can sum to more than one. American Journal of Preventive Medicine, 26 (3), 243-249. (F)

Ruch, G.M., & Stoddard, G.D. (1925). Comparative reliabilities of five types of objective examinations. Journal of Educational Psychology, 15, 89-103. (13)

Rudner, L.M., & Schafer, W.D. (2001). Reliability. ERIC Digest ED458213. (7 pages). (1)

Rulon, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves. Harvard Educational Review, 9, 99-103. (8)

Saupe, J.L. (1966). Selecting items to measure change. Journal of Educational Measurement, 3, 223-228.

Savedra, M.C., Tesler, M.D., et al. (1989). Pain location: Validity and reliability of body outline markings by hospitalized children and adolescents. Research in Nursing & Health, 12, 307-314. (3)

Sawilowsky, S. (2000a). Psychometrics vs. datametrics: Comment on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. Educational and Psychological Measurement, 60, 157-173. (1)

Sawilowsky, S. (2000b). Reliability: Rejoinder to Thompson and Vacha-Haase. Educational and Psychological Measurement, 60, 196-200. (1)

Scheines, R. (2008). Causation, truth, and the law. Brooklyn Law Review, 73 (3), 959-984. (F)

Schmidt, F.L. (1996). Significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1, 115-129. (8)

Schmidt, F.L., & Hunter, J.E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. Psychological Methods, 1, 199-223. (4)

Schmidt, F.L., & Hunter, J.E. (1999). Theory testing and measurement error. Intelligence, 27 (3), 183-198. (2)

Schmidt, F.L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for individual difference constructs. Psychological Methods, 8 (2), 206-224. (2)

Schmidt, F.L., Viswesvaran, C., & Ones, D.S. (2000). Reliability is not validity and validity is not reliability. Personnel Psychology, 53 (4), 3-15. (8)

Schmitt, N. (1996). Uses and abuses of coefficient alpha. Psychological Assessment, 8, 350-353. (8)

Scott, W.A. (1960). Measures of test homogeneity. Educational and Psychological Measurement, 20, 751-757. (8)

Shavelson, R.J., & Webb, N.M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage. (3,13)

Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989). Generalizability theory. American Psychologist, 44, 922-932. (13)

Shoemaker, D.M. (1969). Note on the attenuating effect of zero-variance items on K-R 20. Journal of Educational Measurement, 6, 255-256. (8)

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428. (9)

Siegel, M., & Skeer, M. (2003). Exposure to secondhand smoke and excess lung cancer mortality risk among workers in the "5 B's": Bars, bowling halls, billiard halls, betting establishments, and bingo parlours. Tobacco Control, 12, 333-338. (F)

Siegel, S., & Castellan, N.J. (1988). Nonparametric statistics for the behavioral sciences (2nd. ed.). New York: McGraw-Hill. (2)

Sirotnik, K.A. (1980). Psychometric implications of the unit-of-analysis problem (with examples from the measurement of organizational climate). Journal of Educational Measurement, 17, 245-282. (13)

Smith, C.J., Sears, S.B., et al. (1992). Environmental tobacco smoke: Current assessment and future directions. Toxicologic Pathology, 20 (2), 289-303. (F)

Soliman,S., Pollack, H.A., & Warner, K.E. (2004). Decrease in the prevalence of environmental tobacco smoke exposure in the home during the 1990s in families with children. AmericanJournal of Public Health, 94, 314-320. (F)

Spearman, C. (1904). The proof and measurement of the association between two things. American Journal of Psychology, 15, 72-101. (4,8)

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. American Journal of Psychology, 18, 161-169. (4,8)

Spearman, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 171-195. (4,8)

Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction, and search. (2nd. ed.) Cambridge, MA: TheMIT Press. (F)

Sprott, D.A., & Vogel-Sprott, M.D. (1987). Use of the log odds ratio to assess the reliability of dichotomous questionnaire data. Applied Psychological Measurement, 11, 307-316. (7)

Stallings, W.M., & Gillmore, G.M. (1971). A note on “accuracy” and “precision”. Journal of Educational Measurement, 8, 127-129. (1)

Stanbury, M., Chester, D., Hanna, E.A., & Rosenman, K.D. (2008). How many deaths will it take? A death from asthma associated with work-related environmental tobacco smoke. American Journal of Industrial Medicine, 51, 111-116. (F)

Stanley, J.C. (1962). Analysis-of-variance principles applied to the grading of essay tests. Journal of experimental Education, 30, 279-283. (1)

Stanley, J.C. (1967). General and special formulas for reliability of differences. Journal of Educational Measurement, 4, 249-252. (6)

Stanley, J.C. (1971). Reliability. In R.L. Thorndike (Ed.), Educational Measurement (2nd ed., pp. 356-442). Washington, DC: American Council on Education. (3,6)

Stanley, J.C., & Wang, M.D. (1970). Weighting test items and test-item options: An overview of the analytical and empirical literature. Educational and Psychological Measurement, 30, 21-35. (13)

Stark, M.J., Rohde, K., et al. (2007). The impact of clean indoor air exemptions and preemption policies on the prevalence of a tobacco-specific lung carcinogen among nonsmoking bar and restaurant workers. American Journal of Public Health, 97 (8), 1457-1462. (F)

Steele, M., & Mattox, J.W. (1987). Correlation of arm-span and height in young women of two races. Annals of Human Biology, 14, 445-447. (12)

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation, 9 (4). (8)

Stevens, S.S. (1946). On the theory of scales of measurement. Science, 103, 677-680. (7,10)

Stewart, A.L. (1982). The reliability and validity of self-reported weight and height. Journal of Chronic Diseases, 35, 295-309. (12)

Stine, W.W. (1989). Interobserver relational agreement. Psychological Bulletin, 106 (2), 341-347. (7)

Stranges, S., Bonner, M.R., et al. (2006). Lifetime cumulative exposure to secondhand smoke and risk of myocardial infarction in never smokers. Archives of Internal Medicine, 186, 1961-1967. (F)

Subkoviak, M.J. (1976). Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 13, 265-276. (13)

Subkoviak, M.J. (1978). Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 15, 111-116. (13)

Subkoviak, M.J., & Levin, J.R. (1977). Fallibility of measurement and the power of a statistical test. Journal of Educational Measurement, 14, 47-52. (11)

Suen, H.K. (1987). Agreement, reliability, accuracy, and validity: Toward a clarification. Behavioral Assessment, 10, 343-366. (2)

Suen, H.K. (1990). Principles of test theories. Hillsdale, NJ: Erlbaum. (2,5)

Sutcliffe, J.P. (1958). Error of measurement and the sensitivity of a test of significance. Psychometrika, 23, 9-17. (11)

Sutcliffe, J.P. (1965). A probability model for errors of classification. I. General considerations. Psychometrika, 30, 73-96. (7)

Sutcliffe, J.P. (1980). On the relationship of reliability to statistical power. Psychological Bulletin, 88, 509-515. (11)

Swaen, G., & Amelsvoort, L. (2009). A weight of evidence approach to causal inference. Journal of Clinical Epidemiology, 62, 270-277. (F)

Swaminathan, H., Hambleton, R.K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 11, 263-267. (13)

Swineford, F. (1959). Note on "Tests of the same length do have the same standard error of measurement". Educational and Psychological Measurement, 19, 241-242. (8)

Symonds, P.M. (1928). Factors influencing test reliability. Journal of Educational Psychology, 19, 73-87. (2)

Tam, H.P., & Knapp, T.R. (1997). Some measurement anomalies (what to do and not to do about them). Psychological Testing, 44, 113-121. (4)

Taylor, B.N., & Kuyatt, C.E. (1993). Guidelines for evaluating and expressing uncertainty of NIST measurement results. Technical Note #1297. Gaithersburg, MD: National Institute of Standards and Technology. (5)

ten Berge, J.M.F. (2000). Clarification of Cliff and Caruso (1998). Psychological Methods, 5, 228-229. (13)

Terwilliger, J.S., & Lele, K. (1979). Some relationships among internal consistency, reproducibility, and homogeneity. Journal of Educational Measurement, 16, 101-108. (2)

Thomas, A.R., Hedberg, K., & Fleming, D.W. (2001). Comparison of physician based reporting of tobacco attributable deaths and computer derived estimates of smoking attributable deaths, Oregon, 1989 to 1996. Tobacco Control, 10, 161-164. (F)

Thompson, B. (1999). Five methodology errors in educational research: A pantheon of statistical significance and other faux pas. In B. Thompson (Ed.), Advances in Social Science Methodology (Vol. 5, pp. 23-86). Stamford, CT: JAI Press. (1)

Thompson, B. (2001). Effect sizes, stepwise methods, and other issues: Strong arguments move the field. Journal of Experimental Education, 70, 80-93. (1)

Thompson, B. (2002). Score reliability. Thousand Oaks, CA: Sage. (1)

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. Educational and Psychological Measurement, 60, 174-195. (1)

Thomson, G.H. (1940). Weighting for battery reliability and prediction. British Journal of Psychology, 30, 357-365. (13)

Thorndike, R.L. (1951). Reliability. In E.F. Lindquist (Ed.), Educational Measurement (pp. 560-620). Washington, DC: American Council of Education. (1, 3)

Thouless, R.H. (1939). The effect of errors of measurement on correlation coefficients. British Journal of Psychology, 29, 383-403. (4)

Thun, M.J., Henley, S.J., et al. (2006). Lung cancer death rates in lifelong nonsmokers. Journal of the National Cancer Institute, 98 (10), 691-699. (F)

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. Journal of Educational Psychology, 16, 433-451. (13)

Thurstone, L.L. (1932). The reliability and validity of tests. Ann Arbor, MI: Edwards Brothers, Inc. (3)

Topf, M. (1986). Three estimates of interrater reliability for nominal data. Nursing Research, 35 , 253-255. (7)

Topping, J. (1975). Errors of observation and their treatment (4th ed.). London: Chapman & Hall. (2,3,5)

Traub, R.E. (1967). A note on the reliability of residual change scores. Journal of Educational Measurement, 4, 253-256. (6)

Traub, R.E. (1968). Comment on Glass' response. Journal of Educational Measurement, 5, 343-345. (6)

Traub, R.E. (1994). Reliability for the social sciences: Theory and applications. Thousand Oaks, CA: Sage. (3,7,13)
Traub, R.E. (1997). Classical test theory in historical perspective. Educational Measurement: Issues and Practice, 16 (4), 8-14. (3)

Traub, R.E., & Hambleton, R.K. (1972). The effect of scoring instructions and degree of speededness on the validity and reliability of multiple-choice tests. Educational and Psychological Measurement, 32, 737-758. (13)

Traub, R.E., & Rowley, G.L. (1980). Reliability of test scores and decisions. Applied Psychological Measurement, 4, 517-545. (13)

Trisel, B.A. (2007). What is a premature death? Minerva--An Internet Journal of Philosophy, 11, 54-82. (F)

Tucker, L.R. (1949). A note on the estimation of test reliability by the Kuder-Richardson Formula (20). Psychometrika, 14, 117-119. (8)

Tucker, L.R., Damarin, F., & Messick, S. (1966). A base free measure of change. Psychometrika, 31, 457-473. (6)

U.S. Department of Health and Human Services (2006). The health consequences of involuntary exposure to tobacco smoke. Pittsburgh, PA: U.S. Government Printing Office. (F)

Uter, W., & Pfahlberg, A. (1999). The concept of attributable risk in epidemiological practice. Biometrical Journal, 41 (8), 985-993. (F)

Uter, W., & Pfahlberg, A. (2001). The application of methods to quantify attributable risk in medical practice. Statistical Methods in Medical Research, 10, 231-238. (F)

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20. (1)

van Belle, G. (2002). Statistical rules of thumb. New York: Wiley. (2)

van Belle, G. , & Arnold, A. (2000). Reliability of cognitive tests used in Alzheimer's disease. Statistics in Medicine, 19, 1411-1420. (3)

van der Linden, W.J., & Hambleton, R.K. (1997). Handbook of modern item response theory. New York: Springer.

Van Meter, D.S. (1974). Alternative methods of measuring change: What difference does it make? Political Methodology, 1, 125-140. (6)

van Zyl, J.M., Neudecker, H., & Nel, D.G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. Psychometrika, 65, 271-280. (11)

Verran, J.A., Mark, B.A., & Lamb, G. (1992). Psychometric examination of instruments using aggregated data. Research in Nursing & Health, 15, 237-240. (13)

Viswesvaran, C., Ones, C.J., & Schmidt, F.L. (1996). Comparative analysis of the reliability of job performance ratings. Journal of Applied Psychology, 81, 557-574. (8)

Voss, L.D., Bailey, B.J.R., et al. (1990). The reliability of height measurement. Archives of Disease in Childhood, 65, 1340-1344. (D)

Votaw, D.F. (1948). Testing compound symmetry in a normal multivariate distribution. Annals of Mathematical Statistics, 19, 447-473. (3)

Wacholder, S., Armstrong, B., & Hartge, P. (1993). Validation studies using an alloyed gold standard. American Journal of Epidemiology, 137, 1251-1258. (B)

Wainer, H. (2000). Kelley's paradox. Chance, 11 (1), 47-48. (5)

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores. Applied Measurement in Education, 6, 103-118. (13)

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? Educational Measurement: Issues and Practice, 15 (1), 22-29. (8)

Wakefield, J.A., Jr. (1980). Relationship between two expressions of reliability: Percentage agreement and phi. Educational and Psychological Measurement, 40 (3), 593-597. (7)

Waller, N.G. (2008). Commingled samples: A neglected source of bias in reliability analysis. Applied Psychological Measurement, 32, 210-224. (13)

Wallis, W.A., & Roberts, H.V. (1962). The nature of statistics. New York: The Free Press. (2)

Walter, S.D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. Statistics in Medicine, 17, 101-110. (11)

Webster, H. (1960). A generalization of Kuder-Richardson Reliability Formula 21. Educational and Psychological Measurement, 20, 131-138. (8)

Weiss, D. J., & Shanteau, J. (2003a). The vice of consensus and the virtue of consistency. In C. Smith, J. Shanteau, & P. Johnson (Eds.), Psychological explorations of competent decision making. NY: Cambridge University Press. (B)

Weiss, D. J., & Shanteau, J. (2003b). Empirical assessment of expertise. Human Factors, 45, 104-116. (B)

Wewers, M.E., & Lowe, N.K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. Research in Nursing & Health, 13, 227-236. (7)

Wherry, R.J., & Gaylord, R.H. (1943). The concept of test and item reliability in relation to factor pattern. Psychometrika, 8, 247-264. (8)

Wilcox, R.R. (1978). Estimating true score in the compound binomial error model. Psychometrika, 43, 245-258. (13)

Wiley, D.E., & Wiley, J.A. (1970). The estimation of measurement error in panel data. American Sociological Review, 35, 112-117. (6)

Wiley, D.E., & Wiley, J.A. (1974). A note on correlated errors in repeated measurements. Sociological Methods and Research, 2, 172-188. (6)

Willett, J.B. (1988-1989). Questions and answers in the measurement of change. Review of Research in Education, 15, 345-422. (6)

Williams, R.H., & Zimmerman, D.W. (1977). The reliability of difference scores when errors are correlated. Educational and Psychological Measurement, 37, 679-689. (6)

Williams, R.H., & Zimmerman, D.W. (1984). A critique of Knapp's "The (un)reliability of change scores in counseling research". Measurement and Evaluation in Guidance, 17, 179-182. (6)

Winickoff, J.P., Friebely, J., Tanski, S.E., Sherrod, C., Matt, G.E., Hovell, M.F., & McMillen, R.C. (2009). Beliefs about the health effects of "thirdhand" smoke and home smoking bans. Pediatrics, 123, e74-e79. (F)

Winne, P.H., & Belfry, M.J. (1982). Interpretive problems when correcting for attenuation. Journal of Educational Measurement, 19, 125-134. (4)

Woodruff, D.J. (1990). Conditional standard error of measurement in prediction. Journal of Educational Measurement, 27, 191-208. (5)

Woodruff, D.J., & Feldt, L.S. (1986). Tests of equality of several alpha coefficients when their samples are dependent. Psychometrika, 51, 393-413. (11)

Wright, B.D. (1997). A history of social science measurement. Educational Measurement: Issues and Practices, 15 (4), 33-45, 52. (13)

Wright, B.D., & Masters, G.N. (1982). Rating scale analysis: Rasch measurement. Chicago: MESA Press. (13)

Yen, M., & Lo, L-H. (2002). Examining test-retest reliability: An intra-class correlation approach. Nursing Research, 51 (1), 59-62. (9)

Yi, Q., Wang, P.P., & He, Y. (2008). Reliability analysis for continuous measurements: Equivalence test for agreement. Statistics in Medicine, 27, 2816-2825. (3)

Youngblot, J.M., & Casper, G.R. (1993). Single-item indicators in nursing research. Research in Nursing & Health, 16, 459-465. (7)

Zevallos, J.C., Huang, P., et al. (2004). Usefulness of tobacco check boxes on death certificates: Texas, 1986-1998. American Journal of Public Health, 94, 1610-1613. (F)

Zidek, J.V., Wong, H., et al. (1996). Causality, measurement error and collinearity in epidemiology. Environmetrics, 7, 441-451. (2)

Zimmerman, D.W. (1972). Test reliability and the Kuder-Richardson formulas: Derivation from probability theory. Educational and Psychological Measurement, 32, 939-954. (8)

Zimmerman, D.W. (1994). A note on interpretation of formulas for the reliability of differences. Journal of Educational Measurement, 31, 143-147. (6)

Zimmerman, D.W., Brotohusodo, T.L., & Williams, R.H. (1981). The reliability of sums and differences of test scores. Journal of Experimental Education, 49, 177-186. (6)

Zimmerman, D.W., & Williams, R.H. (1965). Effect of chance success due to guessing on error of measurement in multiple-choice tests. Psychological Reports, 16, 1193-1196. (13)

Zimmerman, D.W., & Williams, R.H. (1966). Interpretation of the standard error of measurement when true scores and error scores on mental tests are not independent. Psychological Reports, 19, 611-617. (5)

Zimmerman, D.W., & Williams, R.H. (1982). Gain scores in research can be highly reliable. Journal of Educational Measurement, 19, 149-154. (6)

Zimmerman, D.W., & Williams, R.H. (1986). Note on the reliability of experimental measures and the power of significance tests. Psychological Bulletin, 100, 123-124. (11)

Zumbo, B.D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.), Advances in Social Science Methodology, 5, 269-304. (6)