Some aspects of classical reliability theory & classical test theory

Brian W. Junker
Department of Statistics
Carnegie Mellon University
Pittsburgh PA 15213
brian@stat.cmu.edu

March 1, 2012

1 Reliability

Reliability is the extent to which the test would produce consistent results if we gave it again under the same circumstances. This seems like a simple idea: why would we give a test if the results were not likely to replicate if we gave the test again? It is also an important idea statistically: when we use tests that are unreliable, statistical hypothesis testing is less powerful, correlations are smaller than they should be, etc.

It turns out that a basic discussion of reliability can be based on a very simple mathematical/statistical model, and for that reason this section is more mathematical than most of the rest of this essay. I hope that the main points of this section are clear even if you "read around" the places that are more mathematical. However, in reading the basic discussion in the next several subsections, please keep in mind:

- The model being discussed, the *classical test theory* model, is so simple as to hardly ever be plausible, and yet, hardly ever falsifiable. Therefore, although it is useful for developing important insights about the role and consequences of reliability, and also for developing quantitative indices of reliability, its conclusions must be treated with a grain of salt. It is a fine first approximation in many settings, but there is almost always a more refined model that will produce a clearer picture of the functioning of the test.

- Classical psychometrics is known famously for two mantras:

  - *Higher reliability is always better!*

– *A test cannot be valid unless it is reliable!*

In reading the next several sections you should begin to see where these mantras come from.

However, you do not have to believe the mantras. I think that in many "high-stakes ranking" settings they make sense. But in many other settings that are not high stakes or may not involve linear ranking they may make less sense. In many exploratory settings fo example, using a test that is face- and content-valid but less reliable, may be the only way to go.

Finally it is worth noting that, just as Messick (1998) argued that validity is not a property of the test but rather a property of the test score that should be reconsidered in each new setting in which that test score will be used, Thompson (2003) has argued that reliability also is a property of the test score, not the test, that should be reconsidered in each new setting in which that test score will be used. This is healthy: educational tests are fragile instruments, and it is prudent to consider whether a test that will be used in a setting other than the one in which it was developed or last used, will continue to have good theoretical (validity) and statistical (reliability) properties.

1.1 Classical Test Theory

Almost all discussions of reliability in testing begin with what is known as *classical test theory (CTT)*, also known as "classical true score theory". CTT is not a (dis-)provable scientific model, rather it is a statistical model for test scores. Gulliksen (1950) attributes the basic form of CTT to Charles Spearman in two 1904 papers. Novick (1966) made CTT safe for modern readers, and a detailed modern treatment is given in Lord and Novick (1968).

The "theory" is easy to state: Let X_{it} be the score that person i receives on a test on occasion t. CTT supposes that

$$X_{it} = T_i + \mathcal{E}_{it} \tag{1}$$

where T_i is the person's "true" score and E_{it} is an error or noise term accounting for the fact that transient influences may force $X_{it} \neq T_i$. For example, you may take the same algebra test two different times ($t = 1$ vs $t = 2$): you may get lucky or unlucky in different ways each time (different $\mathcal{E}_{it}$'s for you); but your underlying proficiency at algebra may not change (same T_i). CTT also assumes

- X is unbiased: $E[X_{it}|T_i] = T_i$, and hence $E[\mathcal{E}_{it}] = 0$;
- $\mathcal{E}$ is uninformative: $\text{Cov}\,(T_i, \mathcal{E}_{it}) = 0$.

Statisticians will recognize Equation 1 as a kind of errors-in-variables, variance-components, or one-way random-effects ANOVA model.

Those interested in measurement should note that the assumptions of unbiasedness and uniformativeness basically mean that the test score X is already *valid* for the true score T. This is a very important point: CTT can say almost nothing *directly* about validity because CTT already assumes

that X is a valid measure of something, namely T (it might not be a valid measure of what you want, but it is a valid measure of something).

The main use of CTT is to gain intuition about the effects of $\mathcal{E}$ in attempting to measure T with X. If we let

$$\sigma_X^2 = \text{Var}(X_{it}), \quad \sigma_T^2 = \text{Var}(T_i), \quad \text{and } \sigma_{\mathcal{E}}^2 = \text{Var}(\mathcal{E}_{it})$$

then

$$\begin{aligned}
\sigma_X^2 &= \text{Var}(X_{it}) \\
&= \text{Var}(T_i + \mathcal{E}_{it}) \\
&= \text{Var}(T_i) + 2\text{Cov}(T_i, \mathcal{E}_{it}) + \text{Var}(\mathcal{E}_i) \\
&= \sigma_T^2 + \sigma_{\mathcal{E}}^2
\end{aligned}$$

because $\text{Cov}(T_i, \mathcal{E}_{it}) = 0$.

A big difference between CTT and the usual one-way random-effects ANOVA problems is that typically we only get to observe one replication per cell: that is, the person only takes the test one time. Thus the model cannot really be fitted (or falsified) with data. Nevertheless, CTT provides a useful framework to think about the decomposition of variability in observed scores into components of variability due to the underlying construct, and due to transient measurement error:

$$\sigma_X^2 = \sigma_T^2 + \sigma_{\mathcal{E}}^2. \tag{2}$$

1.2 Test-Retest Reliability

Reliability is supposed to be about how "repeatable" the results of a test are. So let's consider the same test on two different occasions[1]:

$$\begin{aligned}
X_{i1} &= T_i + \mathcal{E}_{i1} \\
X_{i2} &= T_i + \mathcal{E}_{i2}
\end{aligned}$$

The covariance between the two test scores is

$$\begin{aligned}
\text{Cov}(X_{i1}, X_{i2}) &= \text{Cov}(T_i + E_{i1}, T_i + E_{i2}) \\
&= \text{Cov}(T_i, T_i) + \text{Cov}(T_i, E_{i2}) + \text{Cov}(E_{i1}, T_i) + \text{Cov}(E_{i1}, E_{i2}) \\
&= \sigma_T^2 + 0 + 0 + 0
\end{aligned}$$

[1] So we are carefully controlling the administration of the test on each occasion so that the tests are *strictly parallel*: exactly the same true scores T_i's and exactly the same error variances $\sigma_{\mathcal{E}}^2$ on both occasions, and the errors $\mathcal{E}_{it}$ on one occasion are not informative about the true scores or errors on the other occasion.

Table 1: Typical interpretations of test-retest reliability.

Test-retest Reliability	*Proportion of retest variance explained*	*Typical Interpretation*
0.95	0.90	Excellent
0.90	0.81	Good
0.80	0.64	Moderate
0.70	0.49	Minimal
0.50	0.25	Inadequate

And the correlation between these two test scores is:

$$\begin{aligned} r_{XX} &= \text{Corr}\,(X_{i1}, X_{i2}) \\ &= \frac{\text{Cov}\,(X_{i1}, X_{i2})}{\sqrt{\text{Var}\,(X_{i1})}\,\sqrt{\text{Var}\,(X_{i2})}} \\ &= \frac{\sigma_T^2}{\sqrt{\sigma_X^2}\,\sqrt{\sigma_X^2}} \\ &= \frac{\sigma_T^2}{\sigma_X^2} \end{aligned} \tag{3}$$

The quantity $r_{XX} = \sigma_T^2/\sigma_X^2$ is called the (test-retest) *reliability coefficient*.

Another way to relate X_{i1} to X_{i2} would be to build a linear regression model

$$X_{i2} = b_0 + b_1 X_{i1} + \varepsilon_i$$

The proportion of the variance of X_{i2} explained by X_{i1} in this regression is the *squared correlation*, r_{XX}^2 (e.g. Fox, 1997, pp. 90ff.). For example, Table 1 displays several values for reliability and the corresponding proportion of retest variance explained, along with typical verbal interpretations.

1.3 Standard Error of Measurement

Another immediate application of the reliability coefficient r_{XX} is in computing confidence intervals for T_i from X_i. From Equation 2 we see that

$$\begin{aligned} \sigma_\mathcal{E}^2 &= \sigma_X^2 - \sigma_T^2 \\ &= \sigma_X^2(1 - \sigma_T^2/\sigma_X^2) \\ &= \sigma_X^2(1 - r_{XX}) \end{aligned}$$

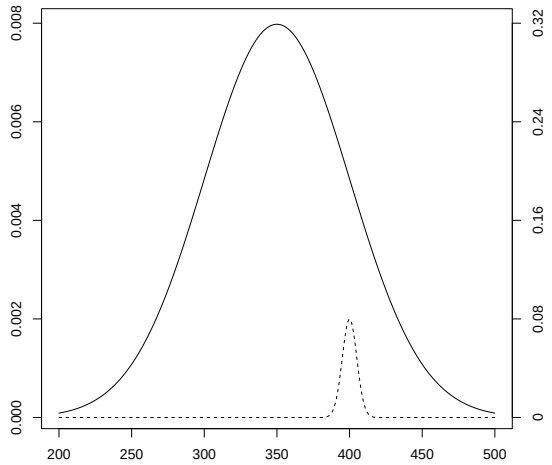

Figure 1: Population SD σ_X indicates spread of scores in the population (solid line); SEM indicates spread of scores measureing the same true score (dashed line).

The *standard error of measurement (SEM)* is the square root of this,

$$SEM = \sigma_X \sqrt{1 - r_{XX}} \tag{4}$$

and thus an approximate 95% confidence interval for T_i would be

$$(X_{it} - 2 \times SEM,\ X_{it} + 2 \times SEM).$$

The SEM should not be confused with the SD of the test scores:

- σ_X = SD of test scores. Describes variation of test scores in the whole population.
- $\sigma_X \sqrt{1 - r_{XX}}$ = SEM of the test. Describes variation of each test score around the corresponding true score.

Figure 1 contrasts the distribution of observed scores in the population (solid line) vs. the distribution of observed scores around measuring the same true score (dashed line). The larger the reliability r_{XX}, the smaller the spread of observed scores around the corresponding true score. However, the spread of scores in the population never gets smaller than σ_T^2.

1.4 Reliability and Statistical Methods

Reliability affects statistical tests, generally by making them less powerful. This generally means you need a larger sample size to detect the same size effect with unreliable measures.

For example, suppose we want to compare the effect of an educational intervention in a treatment group, vs. standard practice in a control group, on a cognitive test. The difference between the mean test scores in each group will be significant if

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{m}S^2 + \frac{1}{n}S^2}} > c_\alpha$$

where c_α is the appropriate level-α cutoff for the t-test. Assuming for simplicity that $m = n$ and inverting this expression we get that the sample size n in each group should be at least

$$n \geq 2\frac{c_\alpha^2 S^2}{(\overline{X} - \overline{Y})^2}$$

Now suppose that we expect to see a difference of $\overline{X} - \overline{Y} = 10$ points in mean test scores between the two groups. We know that

$$S^2 \approx E[S^2] = \sigma_X^2 = \sigma_T^2 / r_{XX}$$

and so the sample size in each group should be, approximately

$$n \geq \frac{c_\alpha^2 \sigma_T^2}{100 r_{XX}}. \tag{5}$$

All other things being equal, the sample size n needed for a significant result is *inversely proportional* to r_{XX}, so again it pays to make r_{XX} as large as possible.

Reliability also attenuates (reduces) correlations between variables. Suppose

$$\left.\begin{array}{rcl} X &=& T^X + \mathcal{E}^X \\ Y &=& T^Y + \mathcal{E}^Y \end{array}\right\} \tag{6}$$

and suppose the $\mathcal{E}$'s are not correlated with each other or with the T's. Then $\text{Cov}(X, Y) = \text{Cov}(T^X, T^Y)$, so that

$$\begin{aligned} \text{Corr}(X, Y) &= \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(T^X, T^Y)}{\sigma_X \sigma_Y} \\ &= \frac{\text{Corr}(T^X, T^Y)\sigma_{T^X}\sigma_{T^Y}}{\sigma_X \sigma_Y} = \text{Corr}(T^X, T^Y)\sqrt{r_{XX} r_{YY}} \end{aligned} \tag{7}$$

This says that

- The maximum of Corr (X, Y) is the correlation of their true scores;
- How close we come to the maximum depends on how high the reliabilities of X and Y are.

This phenomenon is sometimes called the *correlation attenuation* problem because when we use observed scores to compute correlations, they are attenuated toward zero relative to the correlation we would have gotten by using the true scores. A similar phenomenon occurs with regression analyses involving unreliable test scores.

1.5 The Effect of Test Length

The basic formula in Equation 2 says that

$$\sigma_X^2 = \sigma_T^2 + \sigma_\mathcal{E}^2.$$

and we know $r_{XX} = \sigma_T^2/\sigma_X^2$. Usually tests are composed of individual items, and we would like to know what is the effect of lengthening or shortening the test. If we produce a new test $X^{(k)}$ that is k times as long as X, using items with independent errors and the same true score T, the new variance decomposition will be

$$\sigma_{X^{(k)}}^2 = \sigma_T^2 + \sigma_\mathcal{E}^2/k.$$

In that case the reliability of the new test will be

$$\begin{aligned} r_{XX}^{(k)} &= \frac{\sigma_T^2}{\sigma_{X^{(k)}}} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\mathcal{E}^2/k} = \frac{k\sigma_T^2}{k\sigma_T^2 + \sigma_\mathcal{E}^2} = \frac{kr_{XX}}{kr_{XX} + \frac{\sigma_\mathcal{E}^2}{\sigma_X^2}} = \frac{kr_{XX}}{kr_{XX} + 1 - r_{XX}} \\ &= \frac{k \cdot r_{XX}}{1 + (k-1) \cdot r_{XX}} \end{aligned} \tag{8}$$

Equation 8 is called the Spearman-Brown formula. It says for example that a modestly reliable test can be made dramatically more reliable by doubling its length. On the other hand, a test 10 times as long $k = 10$ yields much less than a 10-fold improvement in reliability. A graph of $r_{XX}^{(k)}$ vs. r_{XX} is shown in Figure 2.

1.6 Estimating Reliability

Clearly it would be nice to have an estimate of test-retest reliability. There are several possibilities:

- *Test-retest correlation*. You can give the same test twice to the same group of students, and compute the correlation between the two test scores. This will provide a plausible estimate as long as no indivdidual differnces in (un-)learning takes place between test occasions.

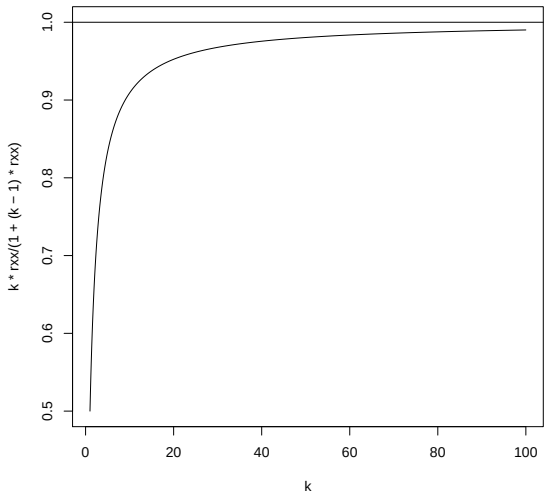

Figure 2: Effect of lengthing the test according to the Spearman-Brown formula. Initial reliability is $r_{XX} = 0.55$.

- *Alternate forms correlation.* If you have two different forms of the same test (say, you have 20 items and you put 10 on one test form and 10 on another test form), you can give both forms to the same people and compute the correlation between the two form scores. This will also provide a plausivle estimate of reliability, though it will likely be lower than the test-retest correlation. Also it depends strongly on the assumption that you were successful at creating two *very* equivalent versions of the test, using different items.

- *Split-Half correlation.* After administering the test you have, split the items up into two equal halves. The correlation between total scores on the two halves is a plausible estimate of reliability. By itself this is probably a lower bound. One possibility to improve it is to apply the Spearman-Brown formula (Equation 8) with $k = 2$, if one believes the two half-tests are really equivalent and have independent error terms.

- *Cronbach's Alpha*. A better estimate of split-half reliability would be to take all the possible splits into equal halves, and average all of the split-half reliabilities. This turns out to be equivalent to Cronbach's (1951) Alpha coefficient:

$$\alpha_C = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \text{Var}\left(X_i\right)}{\sigma_X^2}\right] = \frac{n}{n-1}\left[\frac{\sum \sum_{i \neq j} \text{Cov}\left(X_i, X_j\right)}{\sigma_X^2}\right] \tag{9}$$

for a test of n items, $X_1, \ldots, X_n$. One can show that α is a lower-bound on r_{XX} under mild conditions, (see Novick and Lewis, 1967, for these and other details), and equals r_{XX} under somewhat stronger conditions. With a little algebra, one can also show that

$$\alpha_C \approx \frac{n\overline{r}}{1 + (n-1)\overline{r}}$$

where $\overline{r} = \frac{1}{n(n-1)} \sum \sum_{i \neq j} \text{Corr}\,(X_i, X_j)$ is the average inter-item correlation, which looks quite similar to the Spearman-Brown formula (Equation 8).

Thus we see that an *internal* measure of reliability, Cronbach's alpha, can be used to estimate an *external* measure like test-retest correlation. There are many variations on the idea of using internal consistency to estimate test-retest consistency. Some of the more famous ones are Kuder & Richardson's KR20 (a special case of Equation 9) and KR21 formulae, and more general forms of Equation 9 due to Guttman.

Trochim (2005, 2006) discusses situations in which each of the above estimates is more or less appropriate.

1.7 Other Reliability Issues

1.7.1 Other test reliabilities

There are many reliability or scalability coefficients that account for the particular kind of data one is working with. For example if the items all have dichotomous responses (0 or 1, for incorrect and correct, say) then Mokken (e.g. Mokken, 1997) has suggested using Loevinger's H coefficient as an index of reliability

$$H = \frac{\sum \sum_{i \neq j} \text{Cov}\,(X_i, X_j)}{\sum \sum_{i \neq j} \text{Cov}_{\,max}(X_i, X_j)} \tag{10}$$

where $\text{Cov}_{\,max}(X_i, X_j)$ is the maximum covariance between X_i and X_j that preserves the counts of wrongs and rights for each item separately[2]. This is reminiscent of Cronbach's alpha (Equation 9) and some prefer to use it because alpha can greatly underestimate r_{XX} if the items are dichotomous and the probabilities of success vary greatly from item to item.

Lord (1980) and others have suggested that the basic reliability formula in Equation 3,

$$r_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\text{Var}\,(T)}{\text{Var}\,(X)},$$

[2]This turns out to be easy to compute, since it occurs when the items are *Guttman items*: Whenever the harder question is answered correctly, the easier one is too.

can be generalized beyond the latent variable model in Equation 1 by observing that, since X is unbiased for T, $E[X|T] = T$. Substituting $E[X|T]$ for T above we get

$$r_{XX} = \frac{\text{Var}\,(E[X|T])}{\text{Var}\,(X)},$$

which can be used for any latent variable model at all—even for models in which $E[X|T] \neq T$. This leads to a kind of adaptation of CTT to whatever model one is dealing with, so that all of the formulae above—e.g. those in Section 1.4—apply *approximately* to the new model.

An alternative approach is to take the new latent variable model at face value and develop methodology for hypothesis testing, correlations with other measures, etc. that are exact for that model. This is an approach that has been advocated in more recent theoretical and applied psychometric work, for example, see Mislevy (1991), Fox and Glas (2003), Mariano and Junker (2006), and Schofield et al. (2006). However it is more cumbersome and usually requires collaboration with a quantitative researcher familiar with both modern statistics and measurement.

1.7.2 Inter-rater reliability

A different problem that also goes by the name "reliability" is that of *inter-rater reliability*. We can again build a model like Equation 1,

$$R_{ik} = T_i + \mathcal{E}_{ik} \tag{11}$$

where now R_{ik} is the rating by the k^{th} judge of the i^{th} student's response, T_i is the student's true score as before, and $\mathcal{E}_{ik}$ is rating error, again subject to unbiasedness and uninformativeness assumptions. In this setting we are likely to be able to assign several raters to rate each person's response and so the variance components

$$\sigma^2_R = \sigma^2_T + \sigma^2_{\mathcal{E}}. \tag{12}$$

are usually identifiable and estimable under various practical data collection designs. The resulting reliability coefficient is called the *intraclass correlation*,

$$ICC = \frac{\sigma^2_T}{\sigma^2_R} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_{\mathcal{E}}}. \tag{13}$$

The classic paper on this sort of inter-rater reliability is Shrout & Fleiss (1979). They discuss several designs and several estimators of the ICC. Mariano et al. (2006) reviews several recent approaches that try to address inter-rater reliability in the context of more complex and realistic latent variable models.

Other approaches involve looking at rater agreement. For example if a response can be rated in two or more discrete categories we might look at the *exact agreement*

$$p_{agree} = \frac{\#\{\text{times raters } k_1 \text{ and } k_2 \text{ agree}\}}{\#\{\text{times raters } k_1 \text{ and } k_2 \text{ rated together}\}}$$

or *Cohen's Kappa* (e.g. Landis and Koch, 1977, discuss statistical properties of Kappa, as well as interpretations similar to Table 1), which corrects p_{agree} for chance agreement:

$$\kappa = \frac{p_{agree} - p_{chance}}{1 - p_{chance}}$$

where $p_{chance} = \sum_{\ell} P[\text{ rater } k_1 \text{ rates in category } \ell\ ] \times P[\text{ rater } k_1 \text{ rates in category } \ell\ ]$, or even just the usual product-moment correlation between the raters, r_{k_1,k_2}. While all three of these measures are recommended and/or criticised in various corners of the literature (e.g., Krippendorff and Fleiss, 1978), looking at them together is often useful in identifying raters that need more training, items that are difficult to rate consistently, etc.

1.7.3 Generalizability theory

The basic idea of both Equation 2 and Equation 12 is to decompose the variance of an observed score into a sum of variance components, one for true score and one for error. *Generalizability theory (GT)* extends this decomposition in order to account for several different sources of error together. For example, raters and items might be considered together in a single GT model, so that the effects of increasing the number of test items, as in Equation 8, and increasing the number of raters per item (analogous to Equation 8) can be considered together. Other effects and interactions can also be considered. The classic text is by Cronbach et al. (1972); a recent account from a major methodologist in this area is Brennan (2001); and some recent attempts to embed GT ideas in other psychometric models are presented by Patz et al. (2002).

GT models are formally equivalent to mixed-effects/variance components models, and can be estimated by modern computational Bayes methods such as Markov Chain Monte Carlo (e.g. Mao, Chin and Brennan, 2005), as well as by classical ANOVA and restricted maximum likelihood (REML) methods (e.g. using `PROC VARCOMP` or `PROC MIXED` in SAS, `VARCOMP` or repeated-measures ANOVA in SPSS, or `varcomp` or `lme` in R or Splus).

1.7.4 Item analysis

In addition to evaluating whole tests, it is necessary to construct the tests. For writing cognitive items, Haladyna (1994) and Stiggins (1994) are basic resources. Once the items are written, we need to determine whether they "hang together" as a test. Face validity should play a large role in this process.

In addition several quantitative measures can be used to select items that contribute to high reliability of the total test. This process is called "item analysis", "scale construction", or sometimes just "scaling". Generally speaking we are looking for items X_j that depend on a single, unidimensional[3] latent variable, analogous to T Equation 1. We do this for two reasons: first, such

[3] *Unidimensional* has both a technical meaning and an intuitive meaning here. Technically, a unidimensional vari-

items will generally increase the reliability of the test; and second, a test composed of unidimensional items (those that depend on a unidimensional latent variable) are easier to describe and interpret.

A drawback of this approach is that some important constructs are not unidimensional; for such constructs this approach tends to make tests that are too narrowly focused to be fully valid. For such cases, more complex quantitative methods are required. Fortunately, many constructs can at least be broken down into unidimensional parts—for example, NAEP measures mathematics achievement in five more-unidimensional areas, number properties and operations, measurement, geometry, data analysis and probability, and algebra—so that a unidimensional test can be written for each part.

Here I only mention four of the simplest methods for selecting a set of items that hang together in a unidimensional scale.

- Inspect a table of means and variances for each item, to identify items that are too easy, too hard, or not variable across examinees. Whether or not to keep these items will depend on the *purpose of the test*.
- Inspect a matrix of correlations (or percent agreements, or Cohen's κ's, etc.) between all pairs of items
  - To find coding errors and fix them (e.g. all correlations should be positive)
  - To find groups of items that correlate amongst themselves but not with other items (should these be removed to another subscale?)
- Compute the *point biserial correlation* of each item, that is, the correlation between that item and the total test score. Or compute the *deleted point biserial correlation*, which is the correlation between that item and the total of the *other* items on the test. The point-biserial correlation turns out to be equivalent to the test statistic for a two-sample t-test comparing total score for students who got the item "right" vs students who got it "wrong", so the t-test (on $n-2$ degrees of freedom, if n students took the test) can be used to look for "significant" point-biserials.
- Analogous to the H coefficeint in Equation 10, we can calculate an item-wise H by summing over just on index:
$$H_i = \frac{\sum_j \mathrm{Cov}\,(X_i, X_j)}{\sum_j \mathrm{Cov}_{\,max}(X_i, X_j)}$$
this can be used like the point-biserial to select items into a scale. Sijtsma and Molenaar (2002) discuss an item selection and statistical testing framework.

able is one whose values are numbers on the real line. Intuitively a unidimensional variable corresponds to a construct that is easily and somewhat narrowly characterized, and that you could measure the "amount of" on an ordinal, interval or ratio scale. For example, "Math proficiency" in general is probably not unidimensional, but "proficiency in seventh grade algebra" might well be.

References

Beaton, A.E., & Zwick, R. (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly.* (No. 17-TR-21) Princeton, NJ: National Assessment of Educational Progress/Educational Testing Service.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Borsboom, D. (2006). The attack of the psychometricians. Accepted, *Psychometrika*.

Camilli, G. and Shepard (1994). *MMMS: Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Cronbach, LJ (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* Thousand Oaks, CA: Sage Publications.

Fox, J.P., and Glas, C.A.W. (2003). Bayesian modeling of measurement error in predictor variables. *Psychometrika, 68,* 169–191.

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

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items.* Hillsdale: Lawrence Erlbaum.

Krippendorff, K. and Fleiss, J. L. (1978). Reliability of binary attribute data. *Biometrics, 34,* 142–144.

Kuang, D.C., and Steinberg, L. (2004). *Assessing performance: Investigation of the influence of item context using item response theory methods.* Poster presented at the annual meeting of the Society of Industrial and Organizational Psychology, Chicago, IL.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

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

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

Mao, X., Shin, C. and Brennan, R.L. (2005). *Estimating variability of estimated variance components and related statistics using the MCMC procedure.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal Canada.

Mariano, L. T. and Junker, B. W. (2006). Covariates of the rating process in hierarchical models for multiple ratings of test items. Accepted, *Journal of Educational and Behavioral Statistics.*

Messick, S. (1998). Test validity: A'matter of consequence. *Social Indicators Research, 45,* 35–44.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,*177–196.

Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law, Probability and Risk, 2,* pp. 237–258. Available (March 2006) from `http://ssrn.com/abstract=805060`

Mislevy, R. (2004). Can there be reliability without "reliability"? *Journal of Educational and Behavioral Statistics, 29,* 241–244.

Mislevy, R. J., Almond, R. G. and Lukas, J. F. (2003). *A brief introduction to evidence-centered design.* Unpublished technical report. Available (March 2006) from `http://www.education.umd.edu/EDMS/mislevy/papers/BriefIntroECD.pdf`

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23,* 5–12.

Mokken, R.J. (1997). Nonparametric models for dichotomous responses. In: Hambleton, R.K. and Van der Linden, W.J. (eds). *Handbook of Modern Item Response Theory.* New York-Berlin: Springer-Verlag, pp. 351–367.

National Research Council (NRC). (2001). *Knowing what students know : the science and design of educational assessment*. Committee on the Foundations of Assessment, Center for Education, Division on Behavioral and Social Sciences and Education, National Research Council; James Pellegrino, Naomi Chudowsky, and Robert Glaser, editors. Washington DC: National Academy Press. Available (March 2006) from `http://www.nap.edu/catalog/10019.html`.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1–18.

Novick, M. and Lewis, C. (1967). Coefficient alpha and the reliability of composite measures. *Psychometrika, 32,* 1–13.

Patz, R. J., Junker, B. W., Johnson, M. S. and Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341–384.

Reckase, M. D. (1990). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston MA.

Rothman, R., Slattery, J. B., Vranek, J. L. and Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing.* CSE Technical Report #566. Los Angeles CA: UCLA Center for the Study of Evaluation, National Center for Research on Evaluation, Standards and Student Testing (CRESST). Available (March 2006) from `http://cresst96.cse.ucla.edu/reports/TR566.p`

Russell, M. and Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. *Teachers College Record.* Available (March 2006) from `http://www.tcrecord.org`, ID Number: 10709.

Schoenfeld, A. (2005). Method. To appear in F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning.* New York: MacMillan.

Schofield, L. S., Taylor, L., and Junker, B. W. (2006). The use of cognitive test scores in evaluating black-white wage disparity. Submitted for publication.

Shrout, P.E., and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Sijtsma, K., and Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks: Sage.

Sijtsma, J. and Verweij, A. (1999). Knowledge of solution strategies and IRT modeling of items for transitive reasoning. *Applied Psychological Measurement, 23,* 55–68.

Stiggins, R. (1994). *Student-centered classroom assessment.* New York: Macmillan College Publishing.

Thompson, B. (Ed.) (2003). *Score reliability: Contemporary thinking on reliability issues.* Thousand Oaks: Sage.

Trochim, W. M. K. (2005). *Research methods: the concise knowledge base.* Cincinnati, OH: Atomic Dog Press. Available (March 2006) from `http://www.atomicdog.com`.

Trochim, W. M. K. (2006). *The research methods knowledge base. Third Edition.* Cincinnati, OH: Atomic Dog Press. See also `http://www.socialresearchmethods.net/kb/` for an earlier version.

Yu, C.-H. (2005). *Reliability.* Available (March 2006) from http://seamonkey.ed.asu.edu/~alex/teaching/assessment/reliability.html.