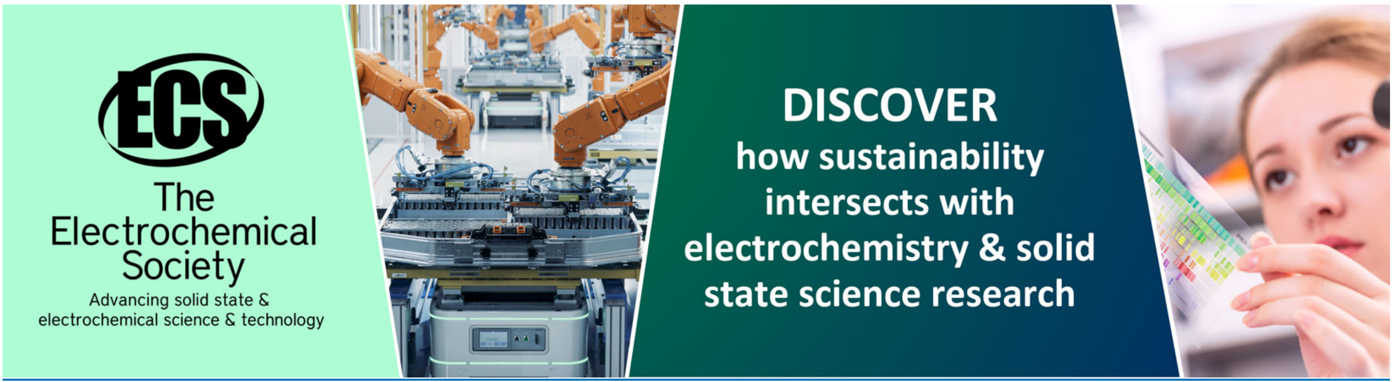
Journal of Physics: Conference Series

Reliability, precision, and measurement in the context of data from ability tests, surveys, and assessments



Reliability, Precision, and Measurement in the Context of Data from Ability Tests, Surveys, and Assessments

W P Fisher Jr[1,2]
[2] Principal and Founder, LivingCapitalMetrics.com
5252 Annunciation St, New Orleans, Louisiana 70115 USA

E-mail: william@livingcapitalmetrics.com

B Elbaum[3]
[3] University of Miami, Florida USA

E-mail: elbaum@miami.edu

A Coulter[4]
[4] Louisiana State University, New Orleans, Louisiana USA

E-mail: acoulter@lsuhsc.edu

Abstract. Reliability coefficients indicate the proportion of total variance attributable to differences among measures separated along a quantitative continuum by a testing, survey, or assessment instrument. Reliability is usually considered to be influenced by both the internal consistency of a data set and the number of items, though textbooks and research papers rarely evaluate the extent to which these factors independently affect the data in question. Probabilistic formulations of the requirements for unidimensional measurement separate consistency from error by modelling individual response processes instead of group-level variation. The utility of this separation is illustrated via analyses of small sets of simulated data, and of subsets of data from a 78-item survey of over 2,500 parents of children with disabilities. Measurement reliability ultimately concerns the structural invariance specified in models requiring sufficient statistics, parameter separation, unidimensionality, and other qualities that historically have made quantification simple, practical, and convenient for end users. The paper concludes with suggestions for a research program aimed at focusing measurement research more on the calibration and wide dissemination of tools applicable to individuals, and less on the statistical study of inter-variable relations in large data sets.

1. Introduction

Reliability is usually conceived and operationalized in terms of a statistical model of inter-variable (item) correlations. The proportion of true variance relative to error variance is expressed as the lower bound of the variance attributable to a common factor tapped by all items, the mean of all possible split-half coefficients, or the correlation that is expected when responses to two random samples of

[1] To whom any correspondence should be addressed.

items are correlated [1,2]. Reliability is then traditionally defined as an estimate of the relation of signal and noise, or of the ratio of the true variance to the error variance [3]:

> The reliability of any set of measurements is logically defined as the proportion of their variance that is true variance…. We think of the total variance of a set of measures as being made up of two sources of variance: true variance and error variance… The true measure is assumed to be the genuine value of whatever is being measured… The error components occur independently and at random (pp. 439-440).

Most reliability coefficients, like Cronbach's alpha, are based in correlational statistical models of group-level information that treats test, survey, or assessment items as separate variables. Alpha incorporates a single standard error estimated from that proportion of the variance not attributable to a common factor. In traditional reliability theory, then, error is the unexplained portion of the variance [2]:

> In his description of alpha Cronbach (1951) proved (1) that alpha is the mean of all possible split-half coefficients, (2) that alpha is the value expected when two random samples of items from a pool like those in the given test are correlated, and (3) that alpha is a lower bound to the proportion of test variance attributable to common factors among the items (pp. 143-4).

Alpha is compromised by its mixing of error and consistency [4]. But in contrast with statistical models of inter-variable relations, measurement models of individual-level response processes [5-9] employ individual-level error estimates, not correlational group-level residual variance estimates. The individual measurement errors are statistically equivalent to sampling confidence intervals. Error and confidence intervals both decline at the same rate with larger numbers of item responses per person, or larger numbers of person responses per item [10].

A reasonable expectation for the measurement standard deviation relative to the error expected on the basis of the number of items and response categories then aids in estimating measurement reliability [11]. Individual-level error terms support more direct estimation of true variance by subtracting the mean square error from the total variance [12, 13]. The availability of individual error terms also supports separate evaluation of sufficiency [14] in terms of modelled expectations [15].

2. The cash value of reliability

Figures 1-7 show that item estimates calibrated on separate samples correlate to about the mean of the scales' reliabilities, and that person estimates measured using different samples of items correlate to about the mean of the measures' reliabilities.

The practical utility of reliability and Rasch separation statistics is that they indicate how many ranges there are in the measurement continuum that are repeatedly reproducible. When reliability is lower than about 0.60, the top measure cannot be confidently distinguished from the bottom one.. Two instruments each measuring the same thing with a 0.60 reliability will produce measures that correlate about 0.60, less well than individual height and weight correlate.

Just as height is not interchangeable with weight, it would be important for practical applications to estimate measures separately for constructs correlating at this level. However, given a 0.60 reliability, a low correlation may be due more to the high proportion of error in the measures than it is to a difference in the constructs measured. In order to distinguish low correlations due to high error from those due to meaningful differences in the constructs measured, correlations should be disattenuated [16]. Low test-retest correlations, correlations of measures from different instruments, or correlations of calibrations from different samples may occur as a function of the reliabilities of the measures or calibrations that are compared, and not because of an inherent lack of association between the constructs measured by different instruments, in different samples, or at different time points. In other words, low correlation may be due more to a high proportion of error in the values compared than to a difference in the constructs measured. Disattenuation does not improve the quality of the measures, and is no substitute for improved measurement precision.

Generally, as reliability increases, so does the number of ranges in the scale that can be distinguished with confidence across samples and/or instruments. Measures with reliabilities of 0.67 will tend to vary within two groups that can be separated with 95% confidence, while those with reliabilities of 0.80 will vary within three groups; of 0.90, four groups; 0.94, five groups; 0.96, six groups; 0.97, seven groups, and so on.

Figures 1-7 illustrate the correlational results (shown in the Table) obtained when (a) different numbers of different items from a single, valid, and reliable instrument [17] are are used to produce multiple measures of the same sample, but with differing reliabilities, and when (b) data from different numbers of respondents in non-overlapping samples are used to produce multiple calibrations of the same items with, again, differing reliabilities. Every pair of measures or pair of calibrations plots linearly, but the widths of the confidence intervals within which the location estimates fall vary markedly. The value of disattenuation (see the Table) can be seen in that, for instance, the wider scatter in the plots and lower r-square values shown in Figures 1-3 and 5-6 are plainly attributable more to low reliabilities and high errors than to a less strong relationship between the sets of measures or calibrations. If the pairs of measures or pairs of calibrations did not correlate well even with large samples or large numbers of items, then the correlation coefficients would be even lower for the data sets with higher errors.

The methodological importance of the point illustrated in Figures 1-7 is widely unappreciated. The calibrations plotted against each other in Figures 1-4 were estimated from completely different samples of respondents from the same population of parents of children with disabilities. Similarly, the measures plotted against each other in Figures 5-7 were estimated by completely different sets of items drawn from a bank of items previously established as measuring the same thing. Given a rough constancy in the amount of variation, error is steadily reduced as items are added, meaning that estimates can be expected to stochastically resonate within a predictably smaller and more precise range [11], increasing G and the number of strata that can be reliably reproduced. The predictive accuracy made possible by reliable and valid measurement has many potential applications in targeting interventions of various kinds in education, psychology, and health care.

3. Discussion

As expected in true score theory [10], reliability estimates tend to systematically increase as the number of items, or sample size, increases. The primary problem with relying on reliability coefficients alone as indications of data consistency hinges on their inability to reveal the location of departures from modeled expectations. Most uses of reliability coefficients take place in contexts in which the model remains unstated and expectations are not formulated or compared with observations. The best that can be done in the absence of a model statement and test of data fit to it is to compare the reliability obtained against that expected on the basis of the number of items and response categories, relative to the observed standard deviation in the scores, expressed in logits [11]. One might then raise questions as to targeting, data consistency, etc. in order to explain larger than expected differences.

Data quality evaluation methods include scatter plots, model fit statistics, latent class analysis of systematic differential person/item functioning, and Principal Components Analysis (PCA). Seeing which individual observations depart the furthest from modelled expectation can illuminate qualitatively meaningful information crucial to correcting data entry errors, identifying sources of bias, separating constructs and populations, and improving the instrument.

The power of the reliability-coefficient-only approach to data quality evaluation is multiplied many times over in a context of a nested series of iterative dialectics in which repeated data analyses explore various hypotheses as to what the construct is, and in which these analyses feed into revisions to the instrument, its administration, and/or the population sampled. Fit analysis should be followed by a return to the drawing board. A theory of the measured construct ultimately provides rigorously precise predictive control over item calibrations, in the manner of the Lexile Framework [18] or developmental theories of hierarchical complexity [19].

4. Conclusion

Measurement is the repetition of a unit amount that maintains its size, within an allowable range of error, no matter which instrument, intended to measure the variable of interest, is used, and no matter who or what relevant person or thing is measured. Measurement is done at the level of the individual, and requires mathematical models of individual response processes. To be meaningful and scientific, instruments must be (1) calibrated to represent additive relations on a number line, and (2) linked together in a network supporting traceability to universally uniform reference standards.

Periodic items recalibration studies show remarkable degrees of stability over time. One study (Bond, 2008) [20] of reading test items used over periods ranging from 7 to 22 years showed 99% of the calibrations changed by less than the minimum detectable difference. Assessments in education and health care capitalize this stability with self-scoring forms providing quantitative results at the point of use, with no need for computerized data analysis [21, 18, 22](Linacre, 1997; Stenner, et al., 2006; Stone, 2002).

As awareness of the capacity to predict calibrations from theory [18, 19, 23, 22](Dawson, 2004; Stenner, et al., 2006; Stenner & Stone, 2003; Stone, 2002) develops, the concept of reliability will evolve. Properly defined and operationalized via a balanced interrelation of theory, data, and instrument (Ackermann, 1985), advanced measurement offers advantages and conveniences that cannot otherwise be obtained. Measurement leadership (Spitzer, 2007) will soon demand calibrated instruments producing reliable measures expressed in meaningful common metrics. When it does, organizations of all kinds will be energized by new possibilities for transformation and empowerment.

8. References

[1] Cronbach L J 1951 *Psychometrika* **16** 297-334

[2] Hattie J 1985 *Applied Psychological Measurement* **9** 139-64

[3] Guilford J P 1965 *Fundamental statistics in psychology and education. 4th Ed* (New York: McGraw-Hill)

[4] Sijtsma K 2009 *Psychometrika* **74** 107-120

[5] Rasch G 1980 *Probabilistic models for some intelligence and attainment tests* (Chicago: University of Chicago Press)

[6] Andrich D 1988 *Rasch models for measurement* (Beverly Hills, CA: Sage Publications)

[7] Wright B D 1977 *Journal of Educational Measurement* **14** 97-116

[8] Fisher W P Jr and Wright B D eds 1994. *Applications of probabilistic conjoint measurement. International Journal of Educational Research* **21** 557-664

[9] Bond T and Fox C 2007 *Applying the Rasch model: Fundamental measurement in the human sciences, 2nd ed* (Mahwah, NJ: Lawrence Erlbaum Associates)

[10] Wainer H and Thissen D 2001 *Test scoring* ed H Wainer and D Thissen (Mahwah, NJ: Lawrence Erlbaum) pp. 23-72

[11] Linacre J M 1993 *Rasch Measurement Transactions* **7** 283-284

[12] Andrich D 1982 *Education Research and Perspectives* **9** 95-104

[13] Wright B D and Masters G N 1982 *Rating scale analysis: Rasch measurement* (Chicago: MESA Press)

[14] Andersen E B 1999 Sufficient statistics in educational measurement *Advances in measurement in educational research and assessment* ed G N Masters and J P Keeves (New York: Pergamon) pp 122-5

[15] Smith R M and Plackner C 2009 *Journal of Applied Measurement* **10** 424-437

[16] Schumacker R E 1996 *Rasch Measurement Transactions* **10** 479

[17] Fisher W P Jr, Elbaum B and Coulter A 2010 *Stakeholder input in scale development, item validation, and standard setting: An example from special education accountability* Unpublished ms

[18] Stenner A J, Burdick H, Sanford E E, and Burdick D S 2006 *Journal of Applied Measurement* **7** 307-22

[19] Dawson T L 2004 *Journal of Adult Development* **11** 71-85
[20] Bond T 2008 *Rasch Measurement Transactions* **22** 1159
[21] Linacre J M 1997 *Physical Medicine and Rehabilitation State of the Art Reviews* **11** 315-324
[22] Stone M 2002 *Knox's cube test - revised* (Wood Dale, IL: Stoelting)
[23] Stenner A J and Stone M 2003 *Rasch Measurement Transactions* **17** 929-30
[24] Ackermann J R 1985 *Data, instruments, and theory: A dialectical approach to understanding science* (Princeton, NJ: Princeton University Press)
[25] Spitzer D 2007 *Transforming performance measurement: Rethinking the way we measure and drive organizational success* (New York: AMACOM)

Acknowledgments. Thanks to Michael Everett and A. Jackson Stenner for their support of this work.

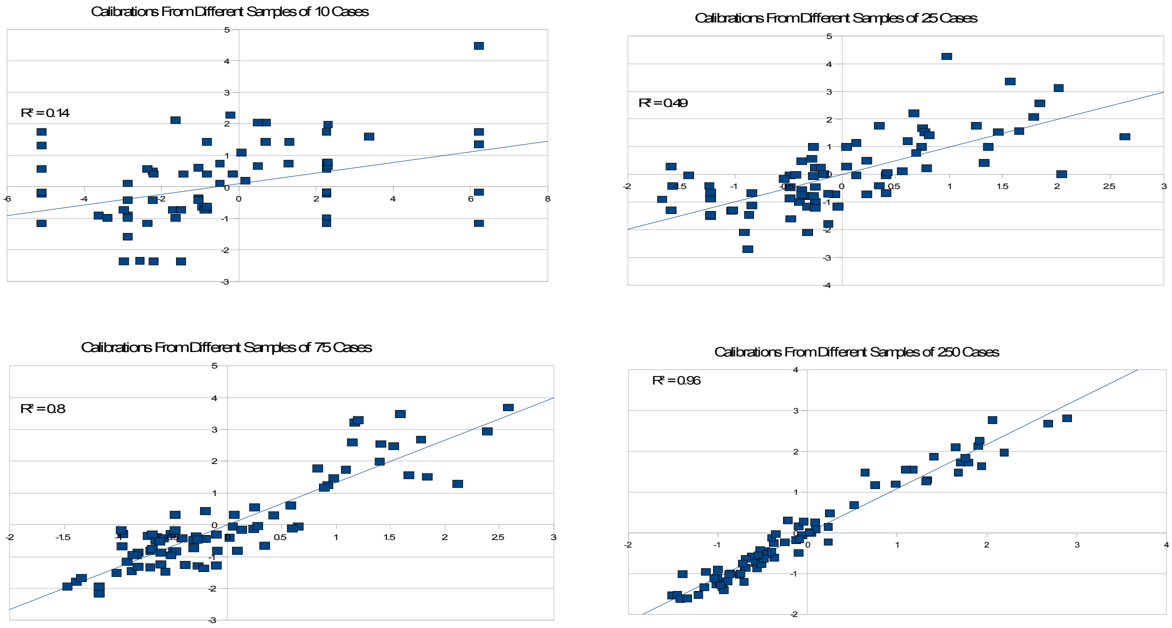
Table. Original and Disattenuated Correlations for Figures 1-7

Figure	Cases Selected and Mean Sample Sizes per Item	Ranges of Winsteps Reliabilities / Separations[2]	Original Correlation	Disattenuated Correlation[3,4]
1	10 / 5.4 / 5.7	0.00-0.66 / 0.00-1.24	0.37	0.56-1.00
2	25 / 14.6 / 15.0	0.59-0.83 / 1.19-2.19	0.70	0.84-1.00
3	75 / 43.9 / 43.9	0.88-0.94 / 2.70-3.83	0.89	0.95-1.00
4	250 / 146.1 / 144.4	0.97-0.98 / 5.54-6.57	0.98	1.00-1.00
Figure	Cases Selected and Mean # of Items per Person	Ranges of Reliabilities / Separations	Original Correlation	Disattenuated Correlation
5	2 / 2.0 / 2.0	0.00-0.59 / 0.00-1.20	0.67	1.00-1.00
6	20 / 12.0 / 11.9	0.88-0.91 / 2.53-3.14	0.80	0.88-0.91
7	52 / 30.9 / 30.2	0.94-0.96 / 3.83-4.70	0.95	0.99-1.00

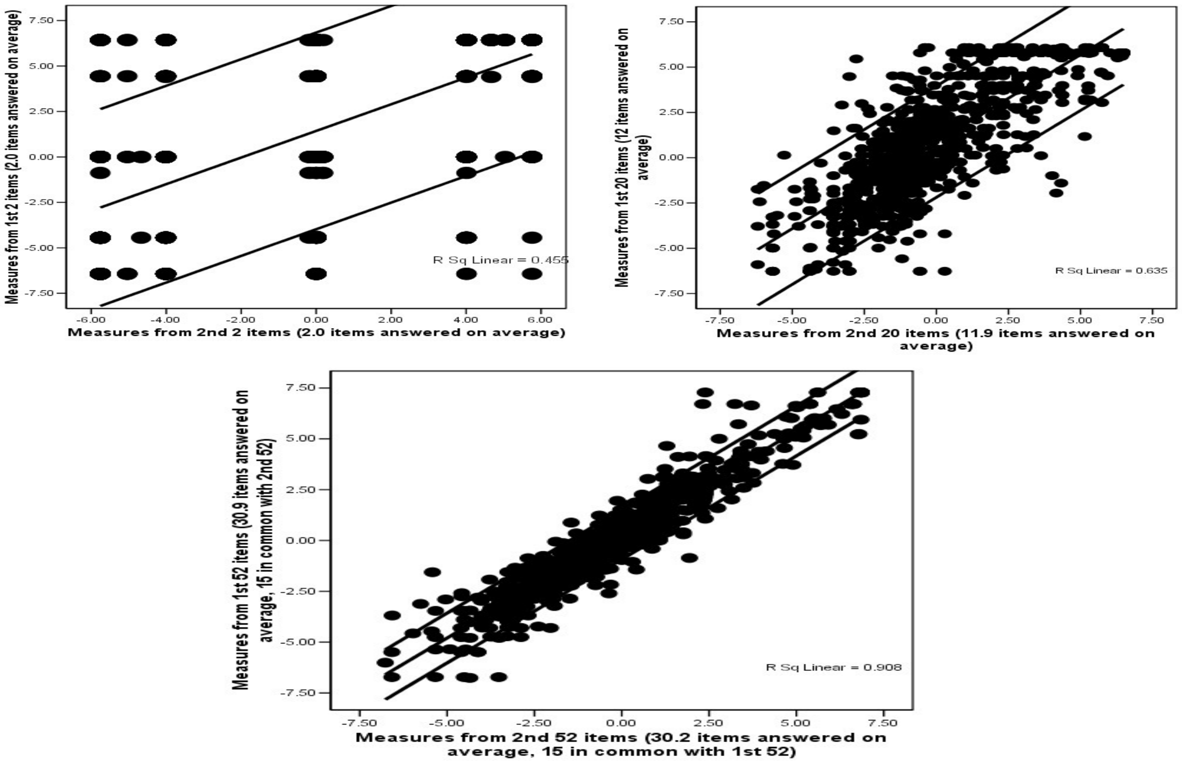
[2] Reliability and separation are estimated separately for the total sample (or all items) and that portion of it (or them) that does not have the minimum or maximum possible scores. Each of these, in turn, are estimated according to the modeled expectations and in a fit-inflated version, for four reliability and four separation values.

[3] The correlations are disattenuated for error by dividing the original correlation by the square root of the product of the two reliabilities.

[4] To illustrate the point as conservatively as possible, pairs of both the minimum and maximum reliabilities in each row were used to estimate the ranges of the disattenuated correlations. Values exceeding 1.00 are shown as 1.00.



Figures 1-4. Scatter plots of calibrations estimated from pairs of different samples of the same size.



Figures 5-7. Scatter plots of measures estimated from pairs of different sets of items.