Resolving the test–retest agreement or reliability dilemma

Hoben Thomas

Abstract
Agreement refers to the degree to which an individual $i$’s test scores match in test–retest settings. Agreement has been thought to be unapproachable with correlational reliability indices. Stable unstable reliability theory extends Spearman’s reliability model and specifies the probability that $i$’s test–retest true scores match. Thus, agreement and reliability are simultaneously addressed. Two examples, one using longitudinal data, illustrate the procedure.

Keywords
Agreement, reliability, test–retest, mixture models

Introduction
In test–retest settings (Berchtold, 2016), reliability is “… the capacity of a test to replicate the same ordering between respondents when measured twice.” More important is agreement. It requires “… the same exact result that each respondent obtains on the two testing occasions” (Berchtold, 2016: 1). That is, test–retest observed scores match for each individual $i$. That agreement is typically unspecified is Berchtold’s principal concern, particularly as agreement relates to longitudinal studies. His example concerns a sample of 678 French-speaking Swiss adolescents responding to the Internet Addiction Test (IAT) test and 5 months later completing a retest. The retest showed a significant mean decrease of 2.77 from the first test. The vexing issue for Berchtold is “… given the small amount of change between the two testing occasions … and the fact that the IAT was evaluated for reliability, but not for agreement, can we really conclude that the IAT is lower” on the second administration than on the first? For Berchtold (2016), the “question is still open (p. 2).”

This article has two goals. The first goal is to resolve the agreement or reliability dilemma by showing that by empirically testing an assumption in Spearman’s random-effects model of correlation (Spearman, 1910; Yule, 1932), the agreement issue can be addressed at the level of each individual $i$. The second goal is to illustrate the theory’s implementation in two examples; the second example explores Berchtold’s IAT longitudinal data.

In 1904, Charles Spearman introduced the notion of reliability in the literature. His name remains associated with the term’s use, at least in the educational and psychological literature. However, no authoritative source has defined reliability as replicating “the same ordering” on both testing occasions (e.g. Gulliksen, 1950; Lord, 1986; Lord and Novick, 1968; Yule, 1932).

However, Berchtold’s definition of agreement is similar to Spearman’s reliability assumption, namely that each $i$’s test–retest true scores match. Another way to represent agreement for Berchtold (2016), Bland and Altman (1986[1986]), and Kottner et al. (2011), among others, is to state that equivalently, pairs of observed scores must lie on a scatter plot’s identity line for agreement to be satisfied, while under Spearman’s reliability theory pairs of true scores (defined below) are assumed to lie on a scatter plot’s identity line. The conceptual difference between defining agreement as based on observed scores and a model-based approach which defines agreement as based on expected scores is crucial, as will be seen.

Department of Psychology, Pennsylvania State University, University Park, PA, USA

Corresponding author:
Hoben Thomas, Department of Psychology, Pennsylvania State University, 121 Moore Building, University Park, PA 16802, USA.
Email: hxt@psu.edu
The extension of Spearman’s model is called stable unstable reliability theory or SURT. Through SURT, empirical tests of Spearman’s assumption are made, and the probability that i’s true scores match is provided. The main ideas of SURT are given below. Details appear elsewhere (Thomas et al., 2012).

**SURT in brief**

**Key points**

- If inference at the i level is desired, modeling must start at the i level.
- The response random variable is a latent structure.
- Matching of test–retest scores (i.e. agreement) is matching of expectations.

Consider modeling i’s IAT scores. Before each measurement is observed, i’s realized values are unknown: That is, they are random variables which are invisible, unobservable quantities. Denote two continuous random variables, Yij, j = 1, 2, where j denotes the first or second IAT test for i. Only realizations of random variables (outcomes) are observable. After responses are rendered, the realized values are for i, for example, Y11 = 11 and Y12 = 9. Random variables are denoted by capitals, and realizations are denoted by lower case letters. The issue is to determine how i’s matching on two trials should be conceptualized. This defines agreement at the i level.

Now consider how Berchtold (2016) and others define agreement: It is matching of each i’s realized values (he uses the term realized) and adds for emphasis “the same exact values” (p. 3). This implies matching occurs when y11 = y12. However, matching on realized values is conceptually fatal for two reasons. First, it ignores the fact that observed realized scores are contaminated with observational or measurement error. So matching on observed scores means matching on perturbed values of continuous variables; such a definition of agreement is conceptually indefensible. Second, matching on realized values is a non-starter if probabilistic structures are to play a role in the evaluation of such matches because to do so requires that continuous random variables take on point values. Y11 = y11 and Y12 = y12 with equality in realizations for agreement to be satisfied: y11 = y12. However, the probability of a continuous random variable taking on any point value is always zero because probability is defined as “area under a curve.”

Consider random variable Y with realizations y and density (probability distribution) f(y). For i’s first IAT of 11, \( P(Y = y = 11) = \int f(y) \, dy \) which in calculus is zero. Consequently, both \( P(Y_{i1} = 11) \) and \( P(Y_{i2} = 9) \) are zero and so is their joint probability.

A conceptual path forward was proposed by Spearman (1910) and developed by Yule (1932). Define \( Y_{ij} \) as follows

\[
Y_{ij} = t_i + E_{ij}, \quad j = 1, 2
\]

This adds a conceptual layer, with \( t_i \) being i’s true score, a fixed individual parameter, and \( E_{ij} \) is a mean zero error random variable. This is Spearman’s “true score error score” latent variable model at the individual i level. Only values of \( y_{ij} \) are observed, while the right-hand side of equation (1) is latent and unobserved. Now each i can have a unique expectation, and it is matching on expectations or true scores that is desired because \( E(Y_{ij}) = t_i \) unperturbed by error (E denotes expectation). Consequently, it is defined following SURT

- i is stable iff \( E(Y_{ij}) = E(Y_{i2}) \).
- i is unstable iff \( E(Y_{ij}) \neq E(Y_{i2}) \).

It is stability that is required for agreement. Instability implies, of course, that i’s true scores are different for each j (which is not represented in the current notation). Spearman’s model assumes that all i are stable. SURT allows this assumption to be evaluated empirically. So, how can i’s stability be specified?

**SURT estimation**

Leave the bivariate correlational setting for the univariate difference score setting using \( y_{i1} - y_{i2} \) to determine the probability i is stable, denoted as wi. Then carry wi back to the correlational setting where w and associated quantities can be informative in multiple ways.

If i is stable, the expected value of the differences is zero: \( E(Y_{i1} - Y_{i2}) = 0 \). Should i be unstable, \( E(Y_{i1} - Y_{i2}) \neq 0 \). So stable i differences will follow a distribution with mean zero, and unstable i differences will follow a distribution with mean non-zero. Determining which distribution i follows is nearly a standard finite mixture distribution problem (Everitt and Hand, 1981). Now assume the error random variables in equation (1), \( E_{ij}, j = 1, 2 \), are independent and normal in distribution, then the distribution of the difference \( Y_{i1} - Y_{i2} \) is normal. This puts the problem into a normal mixture setting where the probability \( w_i \) that i follows the normal distribution with mean zero is part of the standard algorithmic output. The procedure is easy to implement with the R call normalmixEM in the R package mixtools (Benaglia et al., 2009).

**Example 1: peak expiratory flow rate measures**

Bland and Altman (2010 [1986], cited in Berchtold, 2016 and republished in 2010) report test–retest data of 17 subjects measured for peak expiratory flow rate (PEFR; see Table 1). The analysis follows Example 1, 5.1 in Thomas et al. (2012). \( w_i \) estimates are given in Table 1. (Only a very small portion of the output available is reported here.) Figure 1 is a bivariate data plot with circle size surrounding each
data point proportional to \(i\)'s probability of being stable, \(w_i\); the diagonal is the identity line.

Table 1 reveals that individuals 4, 6, and 16 are unstable. These individuals are easily identified in Figure 1 as points surrounded by circles in proportion to \(w_i\), the probability of \(i\) being stable. For example, focus on those stable \(i\) or at least nearly stable. Some individuals are located.

If one's oracle could specify which \(n'\) among a sample of \(n\) are stable, because Spearman's model assumes \(i\) is stable, one would sensibly use just those \(n'\) in computing \(r\). Consider \(r_{w}\) as an estimate of this idealized world.

**Example 2: Berchtold's IAT data**

Usually when reliability is considered, the test–retest interval is such that individuals are assumed to remain stationary between tests. For some adolescents, after 5 months there can be large changes, and so it should not be surprising to see many unstable \(i\). Berchtold's IAT data scatter plot displayed in Figure 2 can be interpreted just as in Figure 1 with points surrounded by circles in proportion to \(w_i\), the probability of \(i\)'s difference scores being from a mean zero stable normal distribution. Many points have no surrounding circle, signaling a zero probability of being stable. The proportion estimated as stable is 0.36, and so most individuals are unstable. The test–retest \(r = 0.71\), while \(r_{w} = 0.94\). If \(w = 0\), \(i\) does not contribute to \(r_{w}\).

The above is premised on the assumption that \(E(Y_{ij} - Y_{ij}) = 0\) for all \(i\) stable. To address Berchtold's concern as to whether, when controlling for stability, the test–retest mean is less than the first test mean, a notational and definitional change is required, so \(i\) can have different true scores at each testing: \(E(Y_{ij}) = t_{ij}, j = 1, 2\). Now redefine stability:

- \(i\) is stable iff \(E(Y_{ij} - \alpha) = E(Y_{ij})\).

The shift parameter \(\alpha = t_{ij} - t_{ij}\). Stable \(i\) difference scores follow a normal distribution with mean \(\alpha\), while unstable \(i\) difference scores follow distributions with other means.
Methodological Innovations

Figure 3 shows the SURT solution with estimate $\alpha = 2.88$. The estimated proportion of $i$ that is stable is 0.62, while $r_w = 0.87$, the SURT reliability.

To determine which of the two solutions represented in Figures 2 and 3 is most appropriate, a penalized log likelihood criterion Bayesian Information Criterion (BIC; McLachlan and Peel, 2000) is used. The larger of the two BIC solutions is declared “the winner.” By this criterion, Figure 3 is the preferred solution.

As noted earlier, Berchtold (2016) reported that the retest was significantly smaller in mean than the first test. His paired $t$-test should be viewed, under SURT, as an unconditional procedure: That is, there is no control for stability or instability of $i$. So there is no possibility of parsing the contribution of the stable or unstable $i$ to the mean difference. SURT is a conditional approach: $\alpha$ is the shift parameter given (i.e. conditioning on) stable $i$. Thus, the SURT solution jointly addresses the issues of stability and shift. This analysis leads to the conclusion that, conditioned on stability or, equivalently, agreement, the retest mean is smaller on retest by an estimated $\alpha = 2.88$.

The SURT analysis also highlights the contrast should agreement be defined as matching on observed scores. IAT scores are integers. Among the 678 respondents, 25 were exact matches: test and retest, $25 / 678 = 0.037$. Under what coherent framework is this proportion to be interpreted? One might plausibly conclude that there is at most negligible agreement. This is quite a different conclusion from the probability model estimates rendered above.

Finally, Figure 3 shows a “fatter” region of circles than Figure 2. This reflects the larger estimated standard deviation of the stable normal distribution associated with the solution of Figure 3.

Discussion

Berchtold (2016) proposes Lin’s (1989) concordance coefficient $\rho_c$ as “an alternative correct method for agreement assessment” (p. 3). The proposal seems peculiar. Lin’s model cannot satisfy Berchtold’s (2016) demands for agreement, namely that matching “is mandatory” at the “individual level for each respondent (p. 3),” because Lin’s model assumes a conventional bivariate random sampling perspective. Lin’s development thus precludes any possible statement concerning individual $i$ agreement. Furthermore, the estimate $\hat{\rho_c}$ is nearly identical to $r$ in many data sets. For example, $\hat{\rho_c} = 0.982$, while $r = 0.983$, and $\hat{\rho_c} = 0.702$, while $r = 0.714$ for Examples 1 and 2, respectively.

If Lin’s model is re-expressed within a conventional latent true score error score formulation while making the standard classical test theory assumptions, the result is

$$\rho_c = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E + (\mu - \mu')^2}$$

with variances $\sigma^2_T$ for true score and $\sigma^2_E$ for error score, and $\mu$ are the true score means on the first and second testings. Under Spearman, these true score means are assumed identical. If $\mu_1 = \mu_2$, Lin’s model is exactly Spearman’s reliability model.

The SURT methodology outlined above provides the tools to determine the probability as to whether $i$ is stable, and SURT does so within a unifying probability framework that blurs the agreement–reliability distinction. It does require some familiarity with R, but the main algorithm can be implemented with a single line of code (Thomas et al., 2012: 209). It appears to perform well with small data sets, as
Example 1 illustrates. SURT does assume normality of the difference scores $y_{i1} - y_{i2}$. However, each $i$ can have its own unique distribution, so random sampling over $i$ is not assumed; furthermore, the parametric requirements on the variances are weaker than the conventional Spearman reliability model. And collectively, SURT’s underlying model assumptions are typically weaker than in many conventional models, including Lin’s (1989).

A Bayes rule classifier is known to be optimal in classification efficiency, given certain conditions; $w_i$ is the corresponding sample version and is recognized as having desirable properties (McLachlan and Peel, 2000: 30–31; Ripley, 1996: 19). However, a practical issue is how large $w_i$ should be to be reasonably assured that $i$ is stable and thus in agreement.

A plausible criterion for deciding $i$ is stable is $w_i > 1/2$. Thus, the estimated probability of $i$ being stable and in agreement is more than one-half. Using this criterion, the first solution for Example 2 (Figure 2) yields 304, while the second solution (Figure 3) yields 535, or 79% of the sample of 678 adolescents are estimated to be stable and in agreement.

Each $w_i$ is variable, however, and consequently, there is uncertainty about the true probability of $i$ being stable. A more demanding criterion is to regard as stable those $i$ which satisfy $w_i - 2\hat{\sigma}_w > 1/2$, where $\hat{\sigma}_w$ is the bootstrap estimated standard deviation of $w_i$ (Efron and Tibshirani, 1993). For the second solution of Example 2, one-third of the sample or 229 satisfy this stringent criterion.

Any generally accepted solution to the reliability or agreement issue is likely to require a model framework based on optimal classification methods, with a model specifically addressing inference at the individual $i$ level. SURT is a framework which does so.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Reference

Note

1. Many thanks to André Berchtold for supplying his IAT data file.

References

Benaglia T, Chauveau D, Hunter DR, et al. (2009) Mixtools: An R package for analyzing finite mixture models. Journal of Statistical Software 32: 1–29.

Berchtold A (2016) Test-retest: Agreement or reliability? Methodological Innovations 9: 1–7.

Bland JM and Altman DG (2010 [1986]) Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet 327: 307–310; International Journal of Nursing Studies 47: 931–936.

Efron B and Tibshirani (1993) An Introduction to the Bootstrap. New York: Chapman & Hall.

Everitt B and Hand DJ (1981) Finite Mixture Distributions. New York: Chapman & Hall.

Gulliksen H (1950) Theory of Mental Tests. New York: John Wiley & Sons.

Kottner J, Audigé L, Brorson S, et al. (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Journal of Clinical Epidemiology 64: 96–106.

Lin LI (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45: 255–268.

Lord FM (1986) Psychological testing theory. In: Kotz S, Johnson NL and Read CB (eds) Encyclopedia of Statistical Sciences (vol. 7). New York: John Wiley & Sons, pp. 343–347.

Lord FM and Novick MR (1968) Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

McLachlan G and Peel D (2000) Finite Mixture Models. New York: John Wiley & Sons.

Ripley BD (1996) Pattern Recognition and Neural Networks. New York: Cambridge University Press.

Spearman C (1904) General intelligence objectively defined and measured. American Journal of Psychology 5: 201–293.

Spearman C (1910) Correlation calculated from faulty data. British Journal of Psychology 3: 271–295.

Thomas H, Lohaus A and Holger D (2012) Stable unstable reliability theory. British Journal of Mathematical and Statistical Psychology 65: 201–221.

Yule GU (1932) An Introduction to the Theory of Statistics (10th edn). London: Griffin Publishing.

Author biography

Hoben Thomas is a Professor of Psychology, Pennsylvania State University.