Title: Assessing test-retest reliability of psychological measures: persistent methodological problems

Running header: Reliability of psychological measures

Victoria K. Aldridge[1], Terence M. Dovey[2], & Angie Wade[1]

[1]Clinical Epidemiology, Nutrition and Biostatistics Section
UCL Great Ormond Street Institute of Child Health,
30 Guilford Street
London
WC1N 1EH, UK

[2]Department of Psychology
Marie Jahoda Building
Brunel University
Kingston Lane
Uxbridge
Middlesex
UB8 3PH, UK

*Author for Correspondence (vicki.aldridge@dmu.ac.uk)
Health and Life Sciences
De Montfort University
Leicester
LE1 9BH, UK
+44 116 2078158

Final published version of this manuscript: https://doi.org/10.1027/1016-9040/a000298

Abstract

Psychological research and clinical practice relies heavily on psychometric testing for measuring psychological constructs that represent symptoms of psychopathology, individual difference characteristics, or cognitive profiles. Test-retest reliability assessment is crucial in the development of psychometric tools, helping to ensure that measurement variation is due to replicable differences between people regardless of time, target behaviour, or user profile. While psychological studies testing the reliability of measurement tools are pervasive in the literature, many still discuss and assess this form of reliability inappropriately with regard to the specified aims of the study or the intended use of the tool. The current paper outlines important factors to consider in test-retest reliability analyses, common errors, and some initial methods for conducting and reporting reliability analyses to avoid such errors. The paper aims to highlight a persistently problematic area in psychological assessment, to illustrate the real-world impact that these problems can have on measurement validity, and to offer relatively simple methods for improving the validity and practical use of reliability statistics.

Key Words: Reliability Analysis, Test-Retest Reliability, Psychometric Testing, Measurement Reliability, Limits of Agreement

Assessing test-retest reliability of psychological measures: persistent methodological problems

Psychometrics is defined by Rust and Golombok (2009) as the science of psychological assessment. Psychological measures assess latent factors such as personality, emotional state, or cognition, via a set of observed variables, and the science of psychometrics is concerned with the quality, validity, reliability, standardization, and removal of bias in such measurement tools (Rust & Golombok, 2009). The vast body of psychological literature utilizing this method of measurement is testament to its value and popularity. However, a great deal of work is required to design and evaluate a new measurement tool to try and ensure that it measures what it intends to, and does so each time it is used. Just as we want to be sure that physical or mechanical tools are giving us the right information every time we use them, we should be equally concerned that the measuring tools we rely upon in research and clinical practice are accurate and dependable. Therefore, consideration of validity and reliability are essential in the development of any new psychological measuring tool.

Validation of psychometric tools ensures that measurements are accurate and meaningful for their target population. Generally, assessments of validity have been well conducted in published psychological research. For instance, multidisciplinary input has long been reported in the development of items and tools (e.g., Bennett & Robinson, 2000; Meyer, Miller, Metzger, & Borkovec, 1990; Steptoe, Pollard, & Wardle, 1995), iterative approaches are usually taken to the refinement of item inclusion, and typically, assessment of both content and performance validities (e.g., construct, criterion-related) are reported (e.g., Garner, Olmstead, & Polivy, 1983; Goodman, 1997, 2001; Pliner & Hobden, 1992).

In contrast, appropriate methods for assessing the reliability of new psychometric measuring tools across time, context, and user (i.e., test-retest reliability), have been more scarcely reported in psychological literature, This is despite the relatively large number of published studies reporting test-retest designs as part of tool development. Because of its predominance and importance in preceding other essential stages in tool development (e.g., scale validation, development of reference/cut-off scores, etc.), the present paper will focus on test-retest reliability, and the statistical problems still observed in this area. While other discussions exist in relation to test-retest methodology, such as the choice of appropriate retest time frames (Chmielewski & Watson, 2009), the focus of this paper will be the analysis of data and presentation of results, rather than study design or data collection. Furthermore, the paper will focus specifically on test-retest of numeric outcomes since these are common outcomes in psychological practice (e.g., test scores, rating scales, etc.) and hence, are often the focus of analysis difficulties. However, some of the key principles also apply to categorical and diagnostic measures.

Many of the topics that will be discussed in the current paper were initially examined by Altman & Bland (1983) within medical research, and issues around types and assessments of reliability have been discussed by numerous authors since that time (e.g., Baumgartner, 2000; Bedard, Martin, Krueger, & Brazil, 2000; Ludbrook, 2002; Streiner, Norman, & Cairney, 2014; Weir, 2005). However, practical changes have been slow to translate into many research domains, including psychology, and this was the rationale for the current paper. The aim of this paper is not to present a comprehensive review of all statistical methods for assessing measurement reliability, or even test-retest reliability specifically, but to discuss some of the fundamental aspects of test-retest reliability analysis that may not be well-understood by researchers undertaking this type of study. The paper will summarise some of the common errors that continue to be observed in published studies, and offer an introduction to relatively simple methods for assessing and reporting test-retest analyses that avoid such errors.

What is test-retest reliability?

Whilst there are many different meanings ascribed to the term ‘reliability’ across scientific disciplines, ‘test-retest’ reliability refers to the systematic examination of consistency, reproducibility, and agreement among two or more measurements of the same individual, using the same tool, under the same conditions (i.e., when we don’t expect the individual being measured to have changed on the given outcome). Test-retest studies help us to understand how dependable our measurement tools are likely to be if they are put into wider use in research and/or clinical practice. When a measurement tool is used on a single occasion, we want to know that it will provide an accurate representation of the patient or participant so that the outcome may be used for practical purposes (e.g., diagnostics, differentiation of individuals or groups). When a measurement tool is used on multiple occasions (e.g., to compare baseline and follow-up) we want to know that the tool will give accurate results on all occasions, so that observed changes in outcome can be attributed to genuine change in the individual, rather than instability in the measurement tool; this is particularly relevant when assessing the efficacy of treatments and interventions. Finally, when a measurement tool is used to assess different groups (e.g., patients receiving different treatments, different characteristics), we want to know that the tool is accurately measuring all individuals so that any group differences may be considered genuine and not an artifact of measurement. Although demonstrating validity is the key to knowing that the right *thing* is being assessed with any given tool, assessing validity is only truly possible once it has been established that a tool is measuring *something* in the same way each time it is used.

In the context of test reliability studies, there are two approaches to understanding the comparability/reliability of test scores – we’ll refer to them in this paper as ‘relative consistency’ and ‘agreement’ – that hold very different definitions of what it means for measurements to be ‘reliable’. Relative consistency, also termed ‘rank-order stability’

(Chmielewski & Watson, 2009), means that the relative position or rank of an individual within a sample is consistent across raters/times, but systematic differences in the raw scores given to individuals by different raters or at different times are unimportant. For example, one assessor may score the first three people in a sample as 100, 105, and 107 for IQ, and the second may score the same three people, at the same time, as 105, 110, and 112. Even though the raw scores given by the two raters are not the same, the difference in rating is consistent across all three participants and they maintain the same rank relative to one another; therefore, the IQ measure would be considered to have relative reliability across raters. In contrast, agreement is concerned with the extent to which the raw observed scores obtained by the measurement tool match (or, agree) between raters or time-points, when measuring the same individual in the absence of any actual change in the outcome being measured.

If the relative ordering of individuals within a given sample is of greater importance or use than the observed differences between individuals (e.g., finishing position in a race) then assessing the relative consistency between measurements may be suitable. However, this is not typically the case when assessing the test-retest reliability of standardized measuring tools such as psychometric questionnaires. In this case, the aim is to try and make objective measurements that are unaffected by the time or place of measurement, or by attributes of the individual making the measurement. Once the tool is applied in practice, we want to be confident that any given measurement is accurate, and that any differences in outcome observed within a study or clinical practice, are due to real changes in an individual, or genuine differences between individuals/groups. Therefore, the purpose of reliability studies in these contexts is to determine the extent to which repeated measurements agree (i.e., are the same), over time, rater, or context (i.e., test-retest), when used to assess the same unchanged individual. In such a case, it is necessary to assess absolute differences in scores, since these provide a direct measure of score stability at an individual level. Aside from the mere presence/absence of stability, absolute score differences also permit the assessment of additional estimates relevant to test-retest reliability, such as the size and homogeneity of differences across sample ranges. References to test-retest reliability are prevalent across psychological questionnaire-based studies, thus acknowledging the perceived importance of accuracy and repeatability in these tools. However, the methods reported to assess 'test-retest' often neglect absolute score differences, and so in many cases they are unsuitable for quantifying the intended form of reliability, and make limit use of the data.

Problems with current methods

Problems in analyses of test-retest reliability most often relate to either unsuitable choice of analysis methods or insufficient reporting of methods. This means, in the first instance, that potentially invalid results are obtained and published (and perhaps trusted in wider practice) and in the second instance, that appraisal and accurate replication of methods and results is precluded. The use of unsuitable statistical methods to assess test-retest reliability may arise from a lack of understanding around different types of reliability, inaccurate understanding of statistical tests/techniques and the results that they produce, or, more pessimistically, as a means for arriving as a desired result where more appropriate or conservative approaches may not (Streiner, 2007). Replication of methodology from published research and resources, some of which may be outdated by the time of use or of poorer quality, can further perpetuate less-than-optimal statistical choices. Inferential statistics are frequently misused in this context and often supersede direct examination and interpretation of the observed differences between measurements. It is very common to see test-retest reliability assessed using bivariate correlation, and non-significant inferential tests of difference, such as paired t-tests, used as evidence of similarity between measurements; neither of which are able to quantify the equality/similarity of repeated scores (Bland & Altman, 1986; Hole, 2014). The following sections summarize the features of these methods that make them unsuitable for assessing agreement-based reliability such as test-retest.

Correlation

Correlation is not agreement. A common misconception is that high correlation between two measurements equates to agreement between them. In reality, quite incongruent paired measurements can produce strong and highly statistically significant correlation coefficients, despite the observed agreement between these measurements being very poor. Parametric correlation coefficients (Pearson's product moment correlations), which are frequently presented in reliability studies, use a -1 to 1 coefficient to quantify how consistently one variable increases (or decreases) relative to another variable increasing, according to how close points lie to any straight line. This can be seen by plotting the measurements against one another and adding a line of best fit. In contrast, agreement in scores means that the 2+ results produced for each individual are the same. To illustrate agreement on a scatter plot the points must lie, not on any straight line, but on the line of equality specifically, where the intercept is 0 and the slope of the line is 1 (Streiner et al., 2014). The difference between correlation and agreement is demonstrated in the data in table 1 taken from a laboratory study of adult food preference. This data shows the ratings given by participants when presented with the same food on two occasions. Despite relative stability of food preferences in adulthood, we see that, even relative to the measurement scale, there are large differences (range 15.7 to 226.7) between ratings given on the two occasions.

Table 1

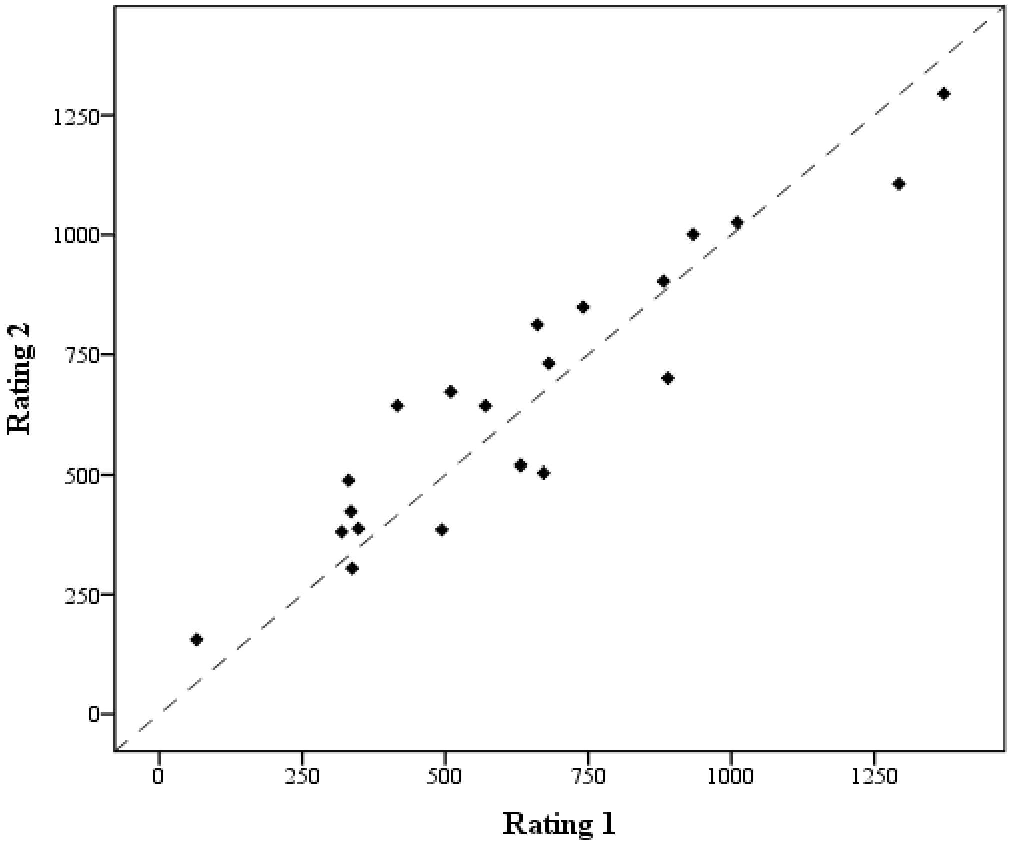
Food ratings on two occasions by n=21 laboratory participants

Person	Rating 1	Rating 2	Difference	Person	Rating 1	Rating 2	Difference
1	1026.10	1010.40	15.70	12	304.83	337.00	-32.17
2	671.93	509.70	162.23	13	812.33	661.00	151.33
3	1001.10	933.20	67.90	14	731.33	681.00	50.33
4	385.40	494.00	-108.60	15	642.73	570.20	72.53
5	503.93	671.80	-167.87	16	519.10	631.80	-112.70
6	848.96	741.00	107.96	17	1295.10	1371.60	-76.50
7	423.76	335.10	88.66	18	387.80	348.00	39.80
8	1106.66	1293.10	-186.44	19	642.96	416.30	226.66
9	381.06	319.00	62.06	20	488.63	330.90	157.73
10	700.36	889.00	-188.64	21	903.03	881.50	21.53
11	156.00	65.50	90.50				

The scatterplot in figure 1 illustrates just how far away paired ratings are from agreement, since very few points lie on or close to the dashed line of equality. Despite this clear disparity in ratings, highlighted in both the plot and the absolute score differences, the Pearson's correlation coefficient for this data is 0.93 ($p<0.001$), which would undoubtedly be reported as a very strong association.

Figure 1

Association between food ratings given at time 1 and time 2, presented against the line of equality (dashed diagonal line).



Correlation conceals systematic bias. Correlation coefficients are standardized statistics that always fall between -1 (perfect negative association) and 1 (perfect positive association). The units and magnitude of the variables being compared are irrelevant in the calculation of the coefficient, and coefficients are not sensitive to mean differences or changes in scores; as such, coefficients will mask systematic biases (the amount that one measurement differs from another) between measurements/measurers. What this means for test-retest reliability is that even very large differences between test and retest values, which may represent significant intra-rater instability or inter-rater difference, will not be detected by correlation analysis if the differences are consistent in a sample. In practice, this means that critical factors affecting measurement reliability such as order effects (practice, boredom, fatigue, etc.) and user interpretation may never be identified. The values in table 2 can be used as an example here; this table presents scores given by two teachers double marking a computer-based exam task and the differences between the scores for each of the 14 students. The table also presents a third set of transformed scores used to exemplify a large difference (bias) in marking. If the way that pairs of measurements increase and decrease relative to one another is constant, the correlation coefficients between measurements will be exactly the same whether there is no bias, a small bias (e.g., around 1.5 points on average), or a very large bias is present (e.g., around 46 points on average). Whilst we are unlikely to see repeated measures differing by such a margin as teachers A and C in real-life data, this more extreme example is used to illustrate an important point. In real world contexts systematic bias can occur if a measurement tool is open to interpretation by the specific user, or where learning and practice effects influence the outcomes of repeated measurements. These are serious flaws for psychometric/psychological assessment tools, which should be unbiased and standardized to permit comparison of measurements within and between samples. Given the substantial negative impact that systematic bias has on agreement-based reliability (values can be far from agreement), methods that conceal and are unaffected by such problems are not appropriate for test-retest reliability analyses.

Although it is uncommon in test-retest reliability, occasionally the measurements that we want to compare have different outcome scales/units, but which denote exactly the same practical result. For example, comparing height measurements between two raters, one of whom uses inches and the other centimeters; or, comparing a total score to a mean or a percentage score. In such cases, the outcomes must be standardized prior to analysis to permit appropriate examination of agreement-based reliability.

Table 2

Comparison of student exam scores showing a small bias (teacher B marks higher than teacher A by ≈1.5 points) and a large bias (teacher C marks higher than teacher A by ≈46 points) in teachers' scores.

Student	Teacher A	Teacher B	Teacher C	B-A Difference	C-A Difference
1	0.0	1.0	4.00	1.00	4.00
2	24.0	24.5	98.00	0.50	74.00
3	19.5	20.0	80.00	0.50	60.50
4	14.0	15.0	60.00	1.00	46.00
5	12.0	13.0	52.00	1.00	40.00
6	0.0	0.0	0.00	0.00	0.00
7	10.5	11.0	44.00	0.50	33.50
8	22.5	22.5	90.00	0.00	67.50
9	6.0	7.5	30.00	1.50	24.00
10	20.5	22.5	90.00	2.00	69.50
11	19.0	22.5	90.00	3.50	71.00
12	8.0	12.0	48.00	4.00	40.00
13	14.5	16.0	64.00	1.50	49.50
14	17.0	20.0	80.00	3.00	63.00
Mean difference				1.43	45.89
Pearson's Correlation between teachers				0.99	0.99

Correlation is influenced by sample variability. A final drawback of correlation analyses is that the strength of a correlation coefficient is influenced by the spread of the measurements in a given sample. More heterogeneous samples will produce stronger correlation coefficients than less heterogeneous samples, in the absence of any disparity in the within-pair measurement differences of each. This means that the resulting correlation coefficient is relative to the sample on which the analysis was based. While absolute differences in coefficients may be relatively small when differences in spread are small, this factor means that it may not be appropriate to directly compare correlation coefficients produced from different samples and populations, and, that coefficients derived from narrow sample ranges may not be representative of the broader population. For example, if you were to compare reliability of growth measurements from a sample of children aged 3-5 years with a sample of children aged 3-10 years, the latter group would be far more variable than the former, so a larger correlation coefficient would be produced for the 3-10 year olds even if agreement in absolute growth measures was the same for both samples. This would also be relevant when comparing reliability estimates from clinical and non-clinical populations, where variation in psychological outcome measures maybe highly disparate between groups. As such the researcher may find that the tool appears more reliable in the non-clinical group than the clinical group (or vice versa), when in actual fact the absolute differences in scores in each group are comparable. It is important to note that this specific issue for test-retest analysis does not arise as a result of narrow or incomparable samples (though these have their own inherent issues if they are unrepresentative), but as a direct result of the use of a relative method (i.e., correlation) to estimate reliability; therefore, it can be overcome by examining absolute differences in scores.

The above issues surrounding correlation analysis also apply to regression analyses when used to assess agreement, since simple regression of one measurement onto another is also based upon association. These problems are particularly hazardous when data are not plotted and examined visually, and reliability is endorsed based on statistical output alone.

An expectation of high agreement between measures may also lead to less rigorous consideration of raw data and statistical results.

Statistical tests of difference

Reliance on traditional statistical testing and p-values can be a hindrance to reliability analysis; "*performing a test of significance for a reliability coefficient is tantamount to committing a type III error – getting the right answer to a question no one is asking"* (Streiner, 2007; Streiner et al., 2014). While there are ongoing debates around the use and/or over-reliance on p-values in research generally, the specific issue in this context is that the null hypotheses against which many such statistical tests are compared are relatively meaningless when assessing reliability. Perhaps the greatest issue relevant to test-retest reliability analysis is the use of hypothesis driven tests of difference, such as the paired t-test. The common fallacy is that, if a test finds no significant difference between measurements then the measurements agree, but this is not the case (Altman & Bland, 1995). Finding a difference to be 'significant' simply means that systematic variability between the measurements (i.e., between raters, conditions, or time-points) outweighs the variability within measurements (i.e., between the individuals in the sample). Therefore, even large differences between repeated measurements, which indicate very poor agreement, can be statistically non-significant if the sample being tested is heterogeneous. The inverse is also true; very similar test-retest scores, which should be seen as demonstrating high reliability, may differ statistically significantly in a homogenous sample.

A related error in reliability analyses is the belief that the average (mean) difference between two or more conditions is adequate to quantify agreement between individual pairs of scores. This error is demonstrated by the data in table 3, which presents another example of laboratory food (pizza) preference ratings (0-20 scale) from 34 participants assessed on two occasions. Table 3 also includes the within-pair differences for scores, the mean score for each time-point, and the mean within-pair difference.

Table 3

Food preference ratings (N=34) for the same food item rated at two time-points, within-pair differences in preference rating (time 1 – time 2), and mean scores.

Case	Time 1	Time 2	Difference	Case	Time 1	Time 2	Difference
1	11.80	11.63	.17	18	14.60	7.91	6.69
2	7.45	7.52	-.07	19	8.85	4.41	4.44
3	9.11	11.80	-2.69	20	9.45	8.32	1.13
4	7.14	9.37	-2.23	21	8.16	9.30	-1.13
5	6.22	8.40	-2.18	22	11.23	12.42	-1.19
6	5.63	8.06	-2.44	23	10.40	9.53	.87
7	6.71	7.88	-1.18	24	13.34	10.09	3.25
8	3.04	2.03	1.01	25	9.02	7.95	1.06
9	9.81	7.62	2.20	26	17.84	8.99	8.85
10	10.21	8.99	1.22	27	14.50	13.41	1.09
11	9.27	9.88	-.62	28	12.18	6.85	5.33
12	4.33	3.78	.55	29	9.60	1.57	8.04
13	4.52	7.10	-2.58	30	13.69	13.54	.15
14	8.90	4.87	4.03	31	7.78	9.00	-1.22
15	4.89	8.67	-3.78	32	12.42	17.16	-4.75
16	7.74	3.86	3.88	33	7.57	10.99	-3.42
17	7.59	7.62	-.02	34	8.54	8.72	-.18
Mean					9.22	8.51	0.71

Relative to the scale of measurement, the absolute differences between ratings are large and variable, ranging from -4.75 to 8.85; and yet, the average within-pair difference is only 0.71. This value suggests far greater similarity in the data than is actually the case. Calculating the mean difference in scores can mask notable disparity between paired measurements. This is particularly true when some scores increase from test to retest and others decrease; whatever the reason for such a pattern in the within-pair differences (e.g., random error of measurement, heteroscedasticity, etc.), this leads to a combination of positive and negative differences that cancel each other out and result in a mean close to zero. When this data is assessed using a paired samples t-test ($t(33) =1.27$, p=0.21) or a Wilcoxon signed rank test to take account of mild skew (Z=258.00, p=0.50), or worse still, an independent samples t-test ($t(66)=0.91$, p=0.37), the difference between measurements is found to be irrefutably non-significant.

Alongside widespread use of correlations and t-tests, a relatively small number of studies in psychology report alternative, more direct methods of analysis for test-retest reliability that utilise absolute differences in scores (e.g., Viglione, Blume-Marcovici, Miller, Giromini, & Meyer, 2012), suggesting gradual improvement in the field. However, where alternative approaches are taken, it can be difficult to determine the validity of the analyses for the given context, or to replicate the methods, because limited methodological detail is reported.

Ways to improve test-retest reliability analysis in psychology

Meet the aims and requirements of test-retest reliability analysis

Typically, test-retest reliability studies are undertaken to see if repeated measurements are ‘similar enough’ for the tool to be considered reliable, which denotes assessment of agreement between raw observed values (i.e., does each individual receive the same value each time they are measured?). The most important outcome for this type of reliability is the size of the differences between related measurements for each individual, rather than whether a difference is seen on average, or whether a 'significant' result is obtained. Traditional hypothesis driven tests assess whether an observed average difference or association is statistically different from zero or no difference/association, rather than indicating how similar/different the observed scores obtained from a tool are. In contrast, suitable methods for analysing test-retest reliability examine the difference(s) between measurements for each case in the sample at an individual level, and assess whether or not the absolute differences between scores obtained by the tool fall within an acceptable range according to the tool's specific clinical, scientific, or practical field of use. Unlike relative consistency, this relies on having an agreed or directly observable unit of measurement for the outcome score. A specific cut-off value (size of difference) up to which measurements may be considered to agree, should be identified and justified by the researcher before viewing the data, to avoid biasing the reliability analyses. Establishing reliability in this way facilitates more in-depth examination of the data (e.g., the size and consistency of differences across a sample) and hence more thorough evaluation of reliability. It also permits the creation and validation of reference values and cut-off scores, for diagnosis and classification and for understanding a single outcome score for an individual; something which is precluded in relative measures of reliability since systematic scoring differences are permissible.

Select suitable methods

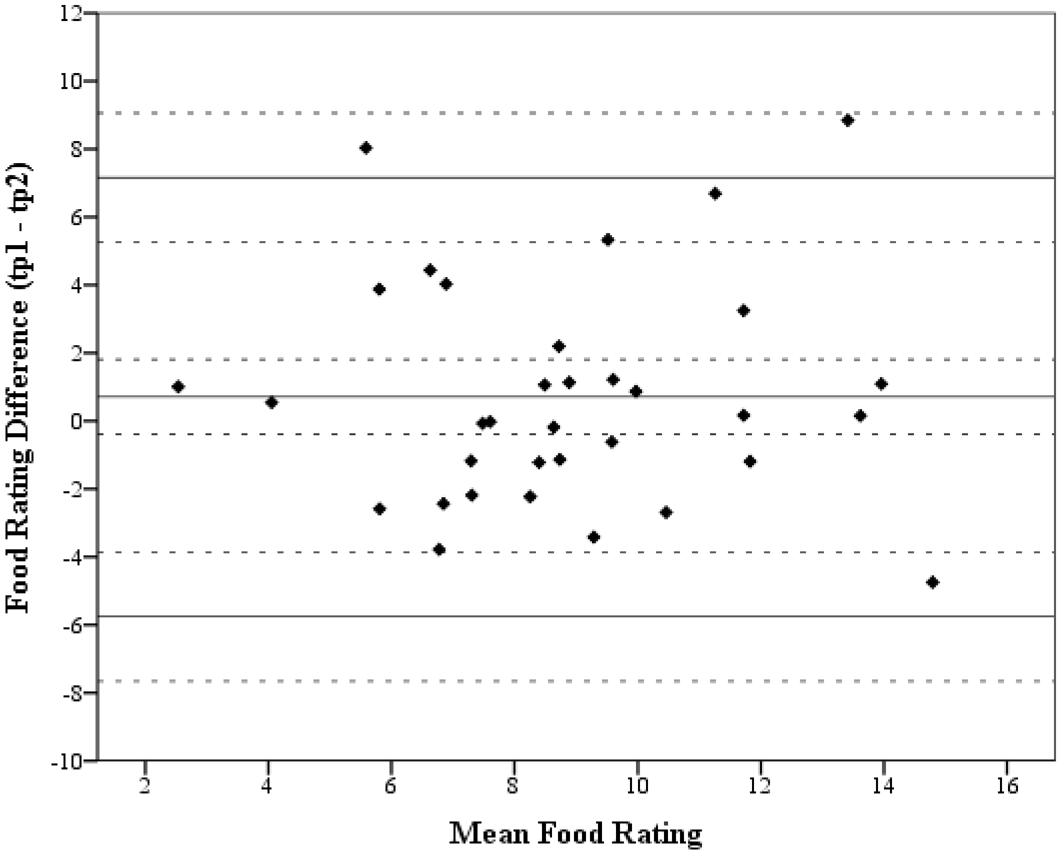
Limits of Agreement. Bland-Altman Limits of Agreement (LOA) (Bland & Altman, 1986) is a statistical method typically used to assess agreement between two repeated numeric measurements (i.e., test-retest scores, or comparison of methods). LOA are based on descriptive statistics for paired data and are typically accompanied by a plot of the data to aid data checking and interpretation. The limits themselves represent the upper and lower boundaries of the middle 95% range of the observed data (within-pair differences), constructed around the mean within-pair difference as mean ± 1.96(SD). For improved interpretation and inference beyond the sample, confidence intervals are also constructed around the upper and lower LOA. Confidence intervals around the LOA will be wider than those around the mean by a factor of 1.71, when samples are not small (Bland & Altman, 1999). Assuming normality of the data, this gives a range of values in which we are 95% confident the population limit should lie. The 'population' is a hypothetical scenario in which all possible measurement differences could be measured, but it provides a practical indication of the variability/precision of measurements that we might expect to see if the tool were implemented widely (e.g., a new clinical or research assessment tool).

An associated Bland-Altman plot sees the average of the two paired measurements plotted on the *x* axis, against the difference between the two measurements on the *y* axis. The plot is used to examine the data distribution and to screen for outliers and heteroscedasticity (where the level of agreement between measurements depends on, or is proportionate to the size of the measurement). When constructing the LOA plot, a horizontal reference line is first added to show the mean within-pair difference; we hope to see this line sitting as close to zero as possible. Points spread evenly either side of zero show random error variability in which measurements do not differ on average. Points lying around any other positive or negative value would indicate systematic bias in the measurements, and if the amount that points vary from the mean line differs across the range of measurements, this suggests that the data are heteroscedastic. Heteroscedasticity may be dealt with via data transformation to permit statistically valid calculation of LOA ((Bland & Altman, 1999). However, it would be essential to try and determine the source of heterogeneity, and to discuss the implications of this data pattern for reliability and wider application of the measurement tool.

If we use the food preference ratings presented in table 3 as an example, we saw previously that the mean within-pair difference for this data was 0.71. Using the standard deviation (3.29) and sample size (*n*=34) we can calculate the standard error (0.56) and a 95% confidence interval for the mean (-0.39, 1.81). The LOA, which represent an interval containing 95% of the observed differences, can be calculated as -5.73 (95% CI -7.64, -3.81) to 7.16 (95% CI 5.24, 9.07); confidence intervals for the LOA are based on a standard error of 0.96 (se mean × 1.71). These key values are added to a Bland-Altman plot (figure 2) to illustrate the extent of agreement, and hence reliability.

Figure 2

Bland-Altman plot showing the mean and limits of agreement (heavy lines) for food preference ratings, and 95% confidence intervals for these estimates (dotted lines).



As we expect, the majority of data points fall within the LOA. If this is not the case, it is likely that the data are skewed and thus, the validity of the LOA is questionable. Figure 2 shows that across the range of observed measurements, data points are randomly scattered around a mean close to zero; this suggests that there is little systematic bias between the two measurements and no obvious data heterogeneity. Negative differences represent a higher score at measure two compared to measure one, while positive differences represent the inverse.

To conclude about agreement, both the LOA plot and statistics should be examined to ascertain how much measurements did (in the observed data) and could (according to the confidence intervals) differ from one another, and if these differences are smaller than a predetermined cut-off for reliability. If there is no systematic bias between measurements (i.e., positive and negative differences are randomly distributed around zero), then either of the limits of agreement (positive or negative) and the confidence interval around that limit can be referenced to conclude about reliability in the wider population. In reality, the mean within-pair difference may deviate a little from zero even from random variation alone, and as shown in our example data, this will lead to an imbalance in the limits of agreement. To make a conservative estimate of reliability, the larger of the two limits should be selected and the confidence interval for this limit used to conclude about agreement. In our example data, the larger of the limits of agreement was 7.16 (95% CI 5.24, 9.07), showing that 95% of the paired measurements in the sample did not differ by more than 7.16 units. In addition, the confidence interval tells us that we can be 95% confident that measurement differences should not exceed 9.07 in the wider population of all measurements.

The disparity between the sample and confidence interval indexes of difference or agreement presented above (7.16 vs. 9.07), illustrates how confidence intervals can alter our conclusions about reliability beyond what is observed in the data, and highlight why it is so important to quantify precision for all estimates. For example, if researchers working with the data had chosen 10 as the maximum difference permitted for this tool to show agreement, we would be confident that our tool was reliable; the chosen cut-off exceeds both the sample and population limits. If instead the cut-off had been 5, we would be quite confident in concluding that our tool was not reliable, since differences observed in the sample and inferred for the population exceed this margin. The most difficult scenario is when the cut-off lies between the two indexes. For example, if the cut-off had been 8, we would have to discuss the implications of our uncertainty around reliability. The observed data do not exceed this value, but reliability is not confidently supported in the context of the wider population. The only way to minimize differences between sample and population estimates is to study large samples, thus reducing the width of confidence intervals around the LOA.

Intraclass Correlation. While there remain frequent problems with reliability analyses in psychology, the use of the Intraclass Correlation Coefficients (ICC) (Shrout & Fleiss, 1979) has been seen in psychological literature for some time (e.g., Angold & Costello, 1995; Egger et al., 2006; Grant et al., 2003; Kernot, Olds, Lewis, & Maher, 2015; March, Sullivan, & Parker, 1999; Silverman, Saavedra, & Pina, 2001). Unlike Pearson's (interclass) correlation, ICC is an acceptable measure of reliability between two or more measurements on the same individual/case. Despite the name and the presence of a coefficient to quantify reliability, ICC is actually based on a ratio of rater, participant, and error sources of measurement variability (derived from ANOVA models). This does mean that ICC coefficients are, like other inferential tests, influenced by sample homogeneity; when variability between measurements is constant, the more alike the sample is, the lower the ICC will be (Bland & Altman, 1990; Lee et al., 2012). Therefore, ICC coefficients derived from samples whose outcome variances differ, such as non-clinical and clinical samples, should not be compared directly. For example, if a depression measure was used in a non-clinical sample we would expect a modest range of scores with many cases scoring close to zero, but this same tool applied to a sample of depressed individuals would likely produce a much greater range of scores. In this case, the clinical sample would obtain a higher ICC coefficient than the more homogenous non-clinical sample, in the absence of any difference in the tool's reliability. This factor does not discredit ICC as a method of reliability analysis, but highlights the importance of evaluating reliability using a representative sample drawn from a relevant population (i.e., in which the tool will be used) (Bland & Altman, 1990). It also emphasizes the need to consider sample variance when interpreting ICC coefficients and differences in reliability observed between samples and populations.

ICC coefficients quantify the extent to which multiple ratings for each individual (within-individual) are statistically similar enough to discriminate between individuals, and should be accompanied by a confidence interval to indicate the precision of the reliability estimate. Most statistical software will also present a p-value for the ICC coefficient. This p-value is obtained by testing sample data against the null hypothesis that measurements within-person are no more alike than between-people (i.e., there is no reliability). In contrast, reliability studies aim to answer the functional question 'are the repeated measurements made using a tool similar enough to be considered reliable'. As such, the p-value provided is, in most cases, of little practical use or relevance.

Though many authors report simply that *'ICC was used'* there are in fact six different ICC types to suit different theoretical and methodological study designs (Atkinson & Nevill, 1998; Shrout & Fleiss, 1979). ICC can be used when a single sample of raters is used to assess every individual in a test sample (type 2 ICC), or when different, randomly selected raters are used across the total sample (type 1 ICC; e.g., when the same individuals cannot feasibly make all measurements across a sample, such as national/multi-center studies). ICC types 1 and 2 quantify agreement. A third ICC type (type 3) is used to assess consistency among a fixed group of raters. Type 3 ICC permits systematic differences between raters, and so represents consistency rather than agreement; therefore, it is only suitable when relative reliability is of primary importance; as discussed earlier in this paper, this is infrequently the case when assessing test-retest reliability for psychometric measures.

For each of the three main ICC types outline above, the coefficient can be calculated in two ways; the first reflects the contributions of each individual rater (e.g., presented as *Single Measures* in SPSS, *Single_raters* in R, and *Individual* in STATA), while the second uses an average of raters (*Average Measures* in SPSS, *Average_raters* in R, or *Average* in STATA). Average options will always result in a larger coefficient, because averaging dilutes the differences across raters/ratings and gives a false inflation of agreement.

Three ICC types and two methods of calculation for each, translates into six different ICC coefficients that could be calculated for any given set of data. However, the coefficients will differ in size, the meaning of ‘reliability’ that they represent, and validity for the particular study. Valid choice of ICC type should be determined by the selection of raters in the particular study, whether or not reliability needs to be generalized to a wider population (e.g., inter-rater reliability generalized to other clinicians using a given measure), and whether consistency or agreement is required. This decision should be clearly outlined in the methods section of research reports (Atkinson & Nevill, 1998; Krebs, 1986).

As an applied example, we can revisit the bias data in table 2. This table presented data from 3 teachers (A, B, and C); teacher B scored on average 1.5 points higher than teacher A, while teaching C scored on average 46 points higher than A. We saw previously that correlation fails to recognize systematic bias and as such the correlations for A with B and A with C were both 0.99 and highly statistically significant. If we now assess this data with ICC type 2 (assuming a sample of teachers were used to assess the random sample of 14 students) to look at agreement, we find that good reliability is demonstrated for teachers A and B who marked similarly (ICC (single measures) = 0.97), and appropriately, very poor reliability is shown for teachers A and C who marked differently (ICC (single measures) =

0.16). When all three teachers are added into the ICC model we see a negligible increase to 0.18, suitably reflecting poor reliability across all three raters. As expected, when ICC type 3 is run, which treats raters as fixed and allows for systematic variability between raters, the result is a considerably higher ICC coefficient of 0.49, which would be higher still if the bias between raters, however large, was consistent. This again highlights the deficiency of assessing consistency rather than agreement for test-retest types of reliability.

An ICC coefficient can also be accompanied by an ICC plot, which sees the sample cases plotted in the *x* axis, outcome scores on the *y* axis, and different point characters used for each rater/rating. ICC plots illustrate the size and nature of observed differences between raters/ratings, and the clustering of scores within person relative to variability across the sample, which aid the practical interpretation of statistical results. For example, figure 3 presents the exam marking data from table 2 for teachers A and B; from this plot we see that teacher A scores consistently lower than teacher B, indicating a small bias, but in most cases the marks are similar. In contrast, figure 4 presents the table 2 data for all three teachers together. This plot clearly shows that teacher C marks much higher than teachers A and B, representing a large positive bias, and hence poor reliability.

Figure 3

Teachers A and B show good agreement (points close together) for student scores; differences within student are small relative to differences between students.

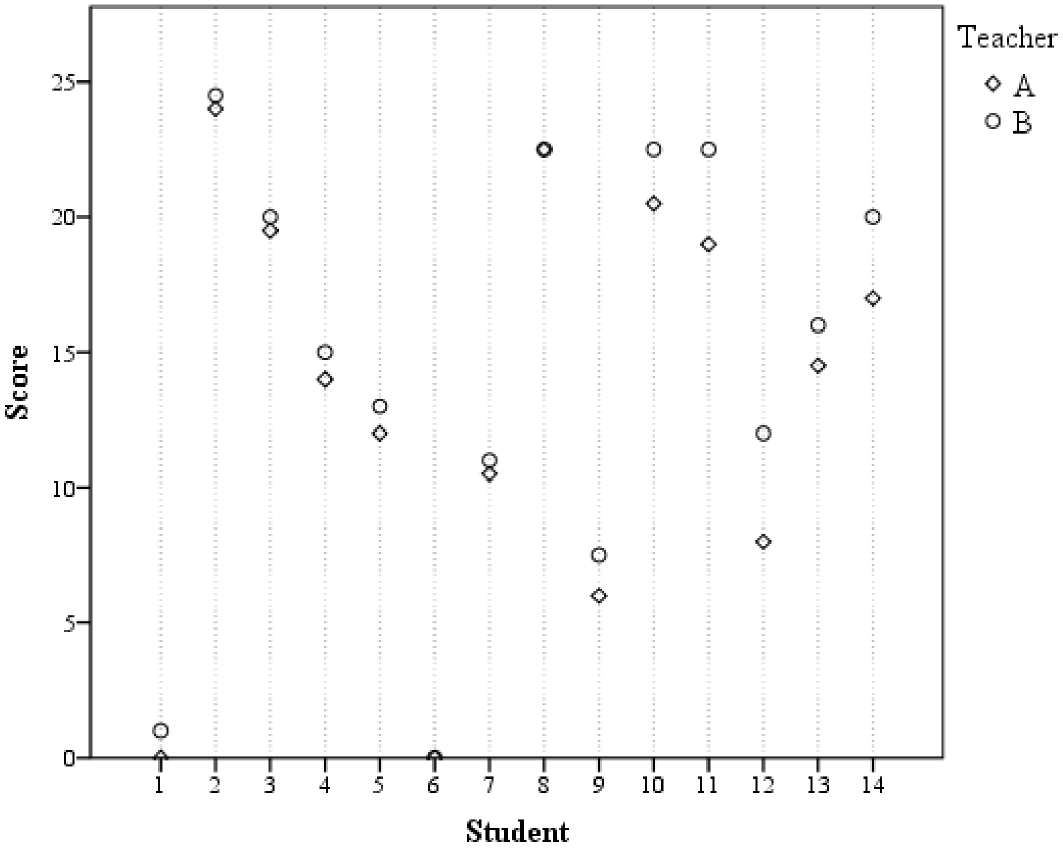
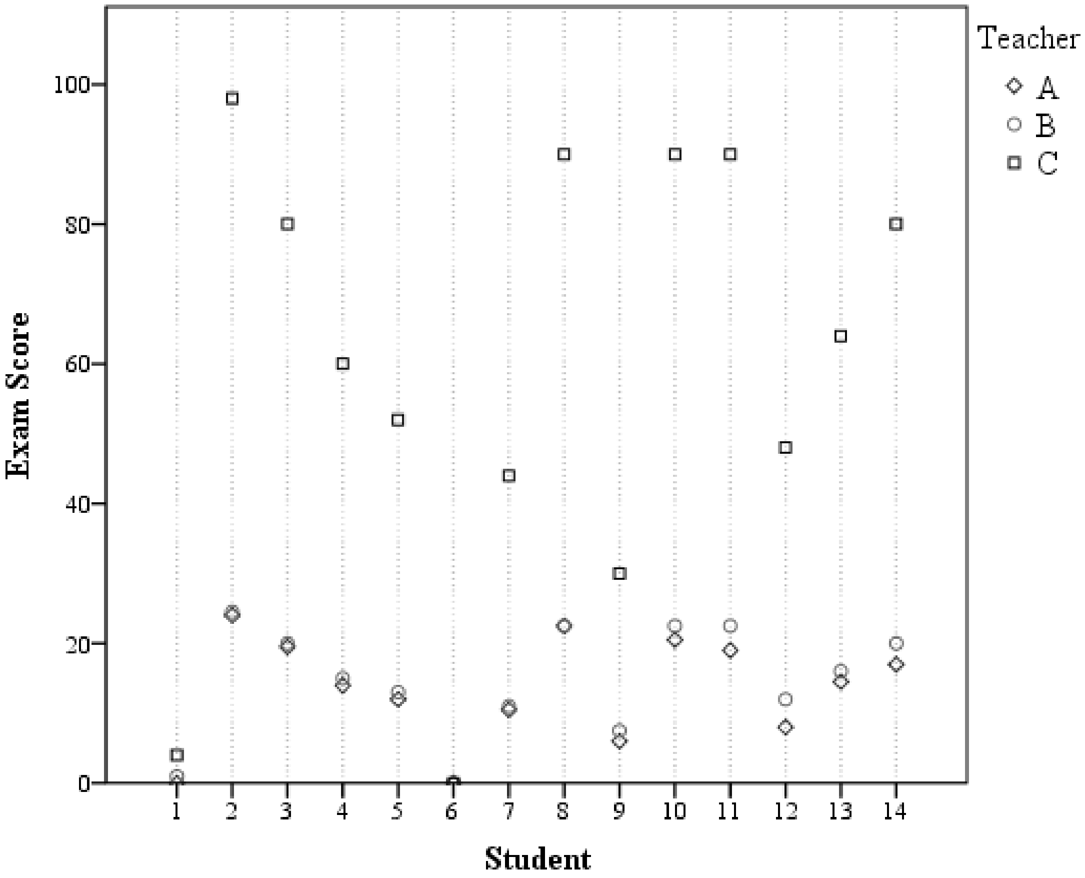


Figure 4

Teachers A, B, and C show very poor agreement (points widely spaced) in student scores. Differences within student are large relative to differences between students.



Improved Reporting

In any study, the aims of the research and the methods used to meet those aims should be clearly outlined; it is insufficient to present vague conclusions verified only by statistical output (e.g., ‘good reliability was shown $r_{(100)}$= 0.85, p=0.006’). The purpose of test-retest reliability studies is to provide evidence that a tool will measure the same thing, in the same way, each time it is used (validity assessment then tells us if it is measuring the right thing). If the methods used to evidence this reliability are not sufficiently explained to validate their use, or the evidence is not presented in the context of a wider population (i.e., no confidence intervals), then the evidence is compromised, or absent altogether. Statements such as ‘ICC was used to assess reliability' are common, despite important differences in ICC models, and the implications of their selection. Such reporting provides no evidence of mindful selection of methods, and may lead the reader to infer that software default settings were used, which vary between packages and may not be appropriate. For example, the default ICC type in IBM SPSS Statistics (version 22) is a two-way mixed effects model for consistency (type 3 ICC). This model is liable to give the highest ICC coefficient of all three main types, but is only appropriate to use when a fixed group of raters is used, and consistent differences between those raters are unimportant. This is contrary to test-retest studies that aim to examine agreement between measurements. It should also be clearly specified and justified when an average of raters is used rather than assessing across individual raters, since this will always inflate the resulting reliability coefficient.

Problems regarding the justification of analytical choices in ICC also extend to correlations and inferential tests of difference. Often, the application of these tests is stated, but neither a rationale for their selection, nor an explanation of how the results demonstrate reliability, are given by the author. These omissions should lead readers to distrust the results, but this is not always the case. Reporting of reliability studies should follow the same recommendations for reporting any research methodology; the information given should be sufficient to allow the reader, in principle, to replicate the study. Complete evidence of reliability, or indeed, unreliability, includes information relevant to the methods and results of the analysis. This should include what will be examined (e.g., agreement between test and retest scores), how this will be assessed (e.g., Bland Altman limits of agreement), and why the method was chosen (e.g., because limits of agreement assesses the extent to which paired measurement in a sample agree). Authors should also clearly document what the results of the assessment indicate about the data and about subsequent use of the tool, relative to practical/clinical parameters and requirements for reliability.

There are some good examples of analyses and reporting within the psychological literature (e.g., Grant et al., 2003; Kernot et al., 2015; Tighe et al., 2015) that demonstrate concise yet informative methodological information. The cited authors are clear about the statistical methods they chose; for example, Tighe and colleagues (2015) reported that they "*calculated the intra-class correlations (ICC) using a two-way mixed effects model for the A, B, and total Alda Scale scores*". Similarly, Grant and colleagues (2003) reported that "*For continuous measures, intraclass correlation coefficients (ICC) are presented as measures of reliability. Since our reliability design assumed that interviewers were randomly drawn from a larger population of interviewers, we used a one-way random effects ANOVA model to derive intraclass correlation coefficients (Shrout and Fleiss, 1979).*" As well as providing important details regarding the specific ICC models applied to their data and, in the case of Grant et al (2003), the justification for this choice, both papers also presented 95% confidence intervals alongside their ICC coefficients. This permits a greater level of interpretation regarding the precision and wider applicability of their results. This information gives the reader a much better indication of what specific statistical procedures were carried out, from which they can better judge the suitability and strength of the resulting evidence.

Extending the principles to other tests of reliability

Although categorical data has not been discussed in the current paper, the key principles of reliability analysis that have been discussed within a test-retest design can be directly translated to these types of outcomes. Firstly, data should be considered at an individual, paired level, in both presentation and analysis. Secondly, analyses should assess the agreement between measurements from multiple conditions, times, or raters. And finally, inferential tests of difference/association, such as chi squared and McNemar's tests for categorical outcomes, should be avoided in favor of specific tests of agreement such as kappa (Cohen, 1960) and its extensions (e.g., weighted kappa, generalized kappa).

The current paper has by no means offered an exhaustive list of potential methods for assessing measurement reliability. Instead, two example analyses have been used to illustrate what an appropriate assessment of agreement-based reliability should comprise. The fundamental messages of the current paper aim to help researchers choose a test or method that actually quantifies reliability, draw conclusions about reliability as directly as possible from the data, and recognize that in most cases a p-value, if given, will provide little practical information about the use or reliability of a measurement tool.

References

Altman, D. G., & Bland, J. M. (1983). Measurement in Medicine - the Analysis of Method Comparison Studies. *Statistician, 32*(3), 307-317. doi:10.2307/2987937

Altman, D. G., & Bland, J. M. (1995). Statistics Notes: Absence of Evidence is not Evidence of Absence. *British Medical Journal, 311*(7003), 485-485. doi:http://dx.doi.org/10.1136/bmj.311.7003.485

Angold, A., & Costello, E. J. (1995). A Test-Retest Reliability Study of Child-Reported Psychiatric-Symptoms and Diagnoses Using the Child and Adolescent Psychiatric-Assessment (Capa-C). *Psychological Medicine, 25*(4), 755-762.

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine, 26*(4), 217-238.

Baumgartner, T. A. (2000). Estimating the Stability Reliability of a Score. *Measurement in Physical Education and Exercise Science, 4*(3), 175-178.

Bedard, M., Martin, N. J., Krueger, P., & Brazil, K. (2000). Assessing reproducibility of data obtained with instruments based on continuous measurements. *Experimental Aging Research, 26*(4), 353-365. doi:Doi 10.1080/036107300750015741

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology, 85*(3), 349-360. doi:10.1037//0021-9010.85.3.349

Bland, J. M., & Altman, D. G. (1986). Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet, 1*(8476), 307-310.

Bland, J. M., & Altman, D. G. (1990). A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Computers in Biology and Medicine, 20*(5), 337-340.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135-160. doi:10.1191/096228099673819272

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology, .97*(1), pp. doi:10.1037/a0015618 19586248

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*(1), 37-46. doi:10.1177/001316446002000104

Egger, H. L., Erkanli, A., Keeler, G., Potts, E., Walter, B. K., & Angold, A. (2006). Test-retest reliability of the Preschool Age Psychiatric Assessment (PAPA). *Journal of the American Academy of Child and Adolescent Psychiatry, 45*(5), 538-549. doi:10.1097/01.chi.0000205705.71194.b8

Garner, D. M., Olmstead, M. P., & Polivy, J. (1983). Development and Validation of a Multidimensional Eating Disorder Inventory for Anorexia-Nervosa and Bulimia. *International Journal of Eating Disorders, 2*(2), 15-34. doi:10.1002/1098-108x(198321)2:2<15::Aid-Eat2260020203>3.0.Co;2-6

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Child Psychology & Psychiatry & Allied Disciplines, 38*(5), 581-586. doi:10.1111/j.1469-7610.1997.tb01545.x 9255702

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry, .40*(11), 1337-1345. doi:10.1097/00004583-200111000-00015

Grant, B. F., Dawson, D. A., Stinson, F. S., Chou, P. S., Kay, W., & Pickering, R. (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*(1), 7-16. doi:10.1016/S0376-8716(03)00070-X

Hole, G. (2014). Eight things you need to know about interpreting correlations. Retrieved from http://www.sussex.ac.uk/Users/grahamh/RM1web/Eight%20things%20you%20need%20to%20know%20about%20interpreting%20correlations.pdf

Kernot, J., Olds, T., Lewis, L. K., & Maher, C. (2015). Test-retest reliability of the English version of the Edinburgh Postnatal Depression Scale. *Archives of Women's Mental Health, .18*(2), pp. doi:10.1007/s00737-014-0461-4 25209355

Krebs, D. E. (1986). Declare your ICC type. *Physical Therapy, 66*(9), 1431-1431.

Lee, K. M., Lee, J., Chung, C. Y., Ahn, S., Sung, K. H., Kim, T. W., . . . Park, M. S. (2012). Pitfalls and important issues in testing reliability using intraclass correlation coefficients in orthopaedic research. *Clinics in Orthopedic Surgery, 4*(2), 149-155. doi:10.4055/cios.2012.4.2.149

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: A critical review. *Clinical and Experimental Pharmacology and Physiology, 29*(7), 527-536. doi:DOI 10.1046/j.1440-1681.2002.03686.x

March, J. S., Sullivan, K., & Parker, J. (1999). Test-retest reliability of the multidimensional anxiety scale for children. *Journal of Anxiety Disorders, 13*(4), 349-358. doi:Doi 10.1016/S0887-6185(99)00009-2

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy, 28*(6), 487-495.

Pliner, P., & Hobden, K. (1992). Development of a Scale to Measure the Trait of Food Neophobia in Humans. *Appetite, 19*(2), 105-120. doi:10.1016/0195-6663(92)90014-W

Rust, J., & Golombok, S. (2009). *Modern psychometrics: The science of psychological assessment., 3rd ed*. New York, NY: Routledge/Taylor & Francis Group; US.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Silverman, W. K., Saavedra, L. M., & Pina, A. A. (2001). Test-retest reliability of anxiety symptoms and diagnoses with the anxiety disorders interview schedule for DSM-IV: Child and parent versions. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(8), 937-944. doi:Doi 10.1097/00004583-200108000-00016

Steptoe, A., Pollard, T. M., & Wardle, J. (1995). Development of a measure of the motives underlying the selection of food: the food choice questionnaire. *Appetite, 25*(3), 267-284. doi:10.1006/appe.1995.0061

Streiner, D. L. (2007). A shortcut to rejection: How not to write the results section of a paper. *Canadian Journal of Psychiatry-Revue Canadienne De Psychiatrie, 52*(6), 385-389.

Streiner, D. L., Norman, G. R., & Cairney, J. (2014). *Health measurement scales : a practical guide to their development and use*. Oxford, United Kingdom: Oxford University Press.

Tighe, S. K., Ritchey, M., Schweizer, B., Goes, F. S., MacKinnon, D., Mondimore, F., . . . Potash, J. B. (2015). Test-retest reliability of a new questionnaire for the retrospective assessment of long-term lithium use in bipolar disorder. *Journal of Affective Disorders, 174*, 589-593. doi:10.1016/j.jad.2014.11.021

Viglione, D. J., Blume-Marcovici, A. C., Miller, H. L., Giromini, L., & Meyer, G. (2012). An inter-rater reliability study for the rorschach performance assessment system. *Journal of Personality Assessment, 94*(6), 607-612. doi:10.1080/00223891.2012.684118

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research, 19*(1), 231-240. doi:Doi 10.1519/15184.1