A Simple Guide to Inter-rater, Intra-rater and Test-retest Reliability for Animal Behaviour Studies

Naomi D. Harvey*[1]

[1] School of Veterinary Medicine and Science, University of Nottingham

*Correspondence: Naomi.Harvey@nottingham.ac.uk

Abstract

This paper outlines the main points to consider when conducting a reliability study in the field of animal behaviour research and describes the relative uses and importance of the different types of reliability assessment: inter-rater, intra-rater and test-retest. Whilst there are no absolute methods under which reliability studies should be analysed or judged, this guide highlights the most common methods for reliability analysis along with recommendations and caveats for how the results should be chosen and interpreted. It is hoped that this guide will serve to improve the quality of, and reporting of, reliability studies in animal behaviour research through aiding both the researchers themselves, and reviewers of their manuscripts.

Keywords: reliability; agreement; temporal consistency; animal behaviour; ratings

Contents

1. INTRODUCTION

In behavioural studies, ethologists routinely need to ascertain how reliable an ethological measurement is between, or within, our pool of raters (scorers/observers). However, identifying the appropriate method (or methods) to use to test whether your score is reliable can be challenging. Often, the answers we seek are hidden inside textbooks or handbooks that we may not have access to in our institutional libraries or that are prohibitively expensive to purchase. Additionally, many papers on the topic of reliability testing have been written specifically for the human medical sector, so may not be as readily applicable to animal behaviour studies. The appropriate statistic to evaluate whether your method is reliable will also depend on the type of data you have and the design of your study, so even those of us familiar with reliability analyses cannot always advise, off the top of our heads, a suitable method for an inquiring student or colleague, or know what should have been the correct protocol when reviewing a paper.

Here, I will describe the different types of rater-reliability and the statistics appropriate for evaluating agreement depending on the type of data and study design using clear and straight-forward language. This paper is intended to serve as a reference guide for students in animal behaviour research, and academics who may be new to reliability testing, to help them to choose the appropriate test for their data. Alternatively, this guide should be of use to academics reviewing animal behaviour studies to help them identify whether the appropriate methods were used or if reliability testing is necessary to recommend. Ideally, this guide should be used to help the researcher consider what data they should collect (and how they will analyse it) prior to beginning a reliability study. What you won’t find in this guide are descriptions of the mathematics behind the various methods for reliability analysis, or in-depth examples of the types of bias that can be present in data. For those who are interested in gaining a deeper understanding of the statistics mentioned here I would refer you to the *Handbook of Inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (Gwet, 2014) or articles in the reference list, of which many (but not all) are open access or available through self-archive repositories.

2. TYPES OF RELIABILITY

The term reliability in animal behaviour studies refers to how consistently something occurs, or conversely, how much impact does measurement error / bias have on the data. There are a number of different types of reliability analysis that can be conducted, each of which will provide you with a different perspective, and will be need to be evidenced under different situations.

In the case of inter-rater (between-rater) reliability, researchers evaluate agreement in how consistently different (usually trained) raters can assign the same score or category to the study subjects. Inter-rater reliability is assessed when multiple individuals score the same set of animals either using video footage or via concurrent scoring during live observations. In some cases, such as when collecting data on a rating scale from owners, for companion animals, or keepers for zoo animals, inter-rater reliability may be assessed by getting multiple people who live with or care for the animal to use the measurement tool within a small window of time (e.g., within a few days of each other).

Intra-rater (within-rater) reliability on the other hand is how consistently the same rater can assign a score or category to the same subjects and is conducted by re-scoring video footage or re-scoring the same animal within a short-enough time frame that the animal should not have changed. Both inter- and intra-rater reliability are integral to designing robust studies of animal behaviour. If your ethogram is poorly defined and there is ambiguity in your behavioural categories, then quality of the data will likely be poor due to high measurement error, which will increase the sample size needed to adequately test your hypothesis (Devine, 2003).

Intra-rater reliability is sometimes confused with the similar yet critically different metric; test-retest reliability. Test-retest reliability is a form of temporal consistency. It also commonly involves the same person scoring the same animal at two or more time points but specifically, test-retest reliability evaluates consistency within the animal as compared to its peers over time. Reliability coefficients for test-retest reliability will by necessity be lower than for intra-rater reliability, because they will include both measurement error within the rater and systematic or context-specific biases within the animal. When being examined in reference to personality (or temperament), test-retest reliability is often evaluated via rank order correlations (such as Spearman's rank or Kendall's Tau-b) to test for inter-individual consistency and can be used to identify behaviours that may be reflective of consistent individual differences (e.g. Harvey, Craigon, Sommerville, et al., 2016). A lack of test-retest reliability reflects a lack of consistent individual differences, meaning that the behaviour measured are highly dependent upon the external, or internal, context. Test-retest reliability should only be investigated once adequate rater reliability has been demonstrated, otherwise it becomes impossible to tell whether poor test-retest performance is due to poor standardisation in scoring, or lack of individual consistency, rendering the findings of such studies uninterpretable.

Both inter-rater and intra-rater reliability can be outcomes to report in a study, but they are also useful metrics for benchmarking the training of your pool of raters, or for refining the definition of your ethogram. For example, inter-rater reliability can be assessed at multiple points during the process of training a new rater until it reaches an acceptable level (e.g. Svartberg & Forkman, 2002). Even with regular training and strict criteria for qualification as a rater, there is often an impact of individual rater on the scores (e.g. Ruefenacht et al., 2002), so perfect agreement is rare and should not be expected.

2.1. Intra-rater reliability: potential impact on interpreting individual scores

It's currently not common practice for people to report the range of the difference (or standard deviation of the difference) when assessing intra-rater reliability in animal behaviour, but when designing a rating tool intended for animal behaviour/welfare assessment, I argue here that it should be. The reason being, that when interpreting differences between rating scores taken over time for individual animals where you compare to a baseline, any actual difference could be magnified (or hidden) by measurement error. It's important to know what magnitude of difference can be expected purely through measurement error to get an idea of how confident you can be in the true magnitude of the difference you detect. As an example, in human clinical medicine, a questionnaire-style assessment tool for recording the severity of atopic eczema reports that for intra-rater reliability, 95% of the second scores fell with 2.6 points of the first (on a 0-28 point scale)(Charman et al., 2004). Using the example from the atopic eczema scale, an increase of 6 points between measurements for a single individual, may for example, indicate a small change as the range of typical measurement error is 2.6 points and the change seen is more than twice that. However, if you knew the typical measurement error was greater, let's say 95% of repeated measurements fell within 5.1 points, then a difference of 6 points may not be enough to be considered biologically or clinically significant. This would be especially applicable for applied situations where animals are regularly assessed by the same people, such as in zoo's or working dog organisations. In such scenarios, it could be suitable to conduct intra-rater reliability testing to discern the magnitude of each raters typical variation, in order to adjust their scores accordingly.

3. CHOOSING A RELIABILITY STATISTIC

Before you can choose your reliability statistic, you need to know what your data will look like. Most studies of animal behaviour in the field of ethology will utilise an ethogram, recording behaviour as durations of states (such as time, or percentage of time, spent resting or foraging) or as frequencies of events (such as short-duration behaviours like scratching & yawning, or by recording states as the number of state bouts). In the case of the applied animal sciences, such as studies of animal welfare or personality, behavioural data may be collected as ratings using questionnaire style assessments. Durations will most often fall into the category of continuous data, whilst questionnaire style ratings may be continuous (e.g., visual analogue scale ratings) or ordinal (e.g., likert scale ratings, although many people treat these as continuous ratings in statistical analyses). In the case of frequency data, how the data should be classed depends on the variance present. In some cases, such as rare behaviour, the data can be so zero-inflated that it is most prudent to convert these to nominal binary data to represent the behaviour as ‘seen’ or ‘not seen’. In other cases, frequency data may be well distributed and could be treated as continuous. The reliability coefficient required for your analysis will depend on what type of data you have collected (Fig. 1).

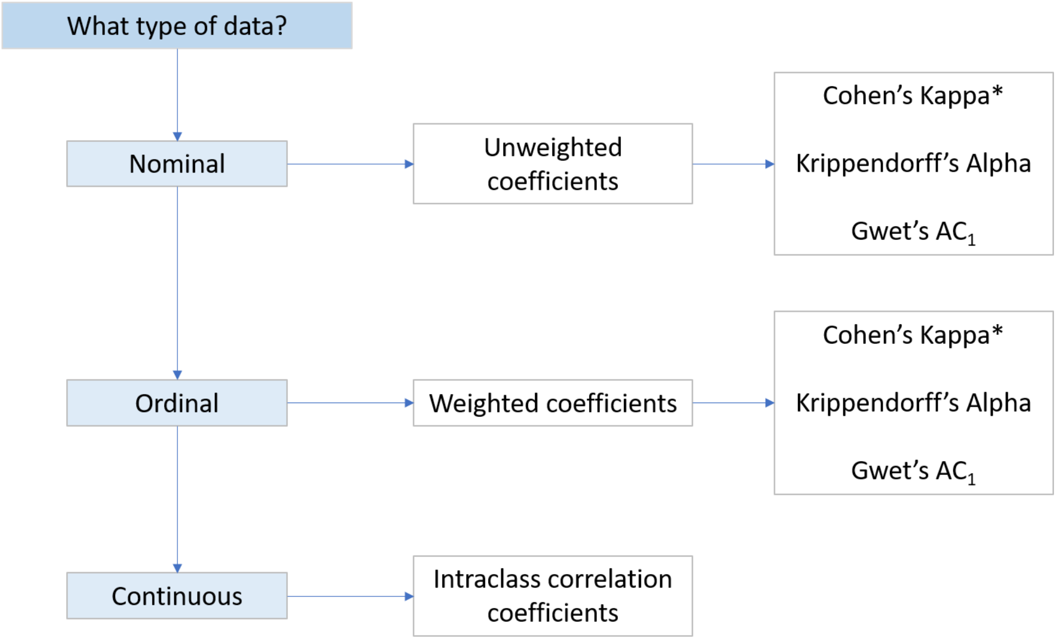

Figure 1. Flow chart depicting selection of suitable inter- and intra-rater reliability coefficients depending on the type of data you need to analyse. *Cohen’s Kappa can only be used where there are just two raters; all other measures are suitable for use with two or more raters.

3.1. RELIABILITY FOR NOMINAL DATA

Binary data is perhaps not so common within animal behaviour research but may be utilised for recording rare events, or for transforming highly zero-skewed data, to indicate whether a behaviour occurred or not during the observation period. The simplest form of reliability testing would be with binary data (two categories) and two raters for inter-rater reliability or the same animals scored twice for intra-rater reliability. In this case, the data can be organised into a simple contingency table (Table 1). It might seem tempting to use percentage agreement to describe such data, however reliability in this case may be overestimated due to the issue of agreement by chance (Gwet, 2014). The appropriate method for evaluating agreement with such data is to calculate an unweighted coefficient that adjusts for chance agreement. The most commonly used coefficient is Cohen's Kappa, which is the percentage of agreement adjusted for chance (Cohen, 1960). Using the example provide in Table 1, the raters agreed on 74% of the observations (calculated as (30+23)/72) and adjusting for chance agreement using Cohen's Kappa gives a Kappa statistic of 0.64.

Table 1. Hypothetical example of a contingency table for a two-rater reliability study where a behaviour was recorded as occurring (1) or not occurring (0) in an observation period.

Rater B	**Rater A**		Total
	0	1	
0	30	11	41
1	8	23	31
Total	38	34	72

Kappa coefficient's range between -1 and 1, with 1 indicating perfect agreement, 0 indicating chance level agreement and -1 indicating perfect disagreement (Viera & Garrett, 2005). Judgement for what level of agreement should be accepted needs to be tailored to the specific context under which your measurements will be applied (McHugh, 2012). The general rule of thumb suggested by Cohen (Cohen, 1960) is that kappa coefficients of agreement can be considered:

- 0.01-0.20 none to slight
- 0.21-0.40 fair
- 0.41-0.60 moderate
- 0.61-0.80 good/substantial
- 0.81-1.00 excellent

These thresholds are an arbitrary guide and should not be treated as definitive. If the consequences of your measurements are serious, such as selection for breeding programmes based upon behavioural indices, then you may wish to consider only basing your decisions only upon indices with 'excellent' reliability of >0.80. If agreement is weak (i.e., <0.60), it would minimise measurement error to use multiple observers if feasible and proceed with your analysis using an average of their scores rather than relying on data from a single rater.

Cohen's Kappa is easy to calculate in most statistical software packages; however, it can only be used in situations with two raters. There are various similar alternatives to Cohen's unweighted Kappa, all of which can be used with situations where there are 3 or more raters, such as Krippendorf's alpha (Hayes & Krippendorff, 2007) (which can be used with less restrictions than Cohen's kappa), Fleiss' Kappa (Fleiss, 1971) (minor differences from Krippendorff's alpha), and Gwet's AC_1 (Gwet, 2008) (described as more paradox resistant than alternatives).

3.2. Reliability for ordinal data

When categories can be ordered, from weak to strong or low to high, this is considered to be ordinal data; a common example of such data would be that collected using a 5-point Likert scale. The basic Kappa coefficient described previously is not suitable for analysing such data, as it does not take into account the magnitude of the difference between data points and considers any disagreement as complete disagreement (Gwet, 2014). When data is ordinal however, raters may assign similar scores, which still amounts to some level of agreement as in the case of Raters 1 and 2 in Fig. 2.

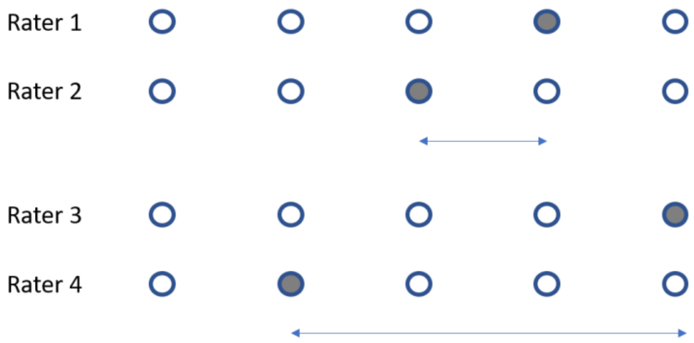

Figure 2. Example of a 5-point Likert scale showing differing extents of disagreement. Raters 1 and 2 have a single point gap in their assigned scores, whilst Raters 3 and 4 have a four-point gap, which ostensibly indicates a much greater level of disagreement.

To overcome this limitation of Kappa coefficients for analysing ordinal data, Cohen proposed a weighted version of his Kappa coefficient (K_W) (Cohen, 1968). Whilst not so commonly found in standard statistical packages, Krippendorf's alpha is a very flexible reliability coefficient, that has fewer limitations than Cohen's Kappa (for example, it can be used with more than two raters) and can also be extended to evaluate ordinal data (Hayes & Krippendorff, 2007). There is a lot of overlap in the use of these measures for analysing ordinal data, with the use of standard correlation coefficients. The weighted Kappa for instance is in some circumstances directly equivalent to a Pearsons *rho* (Cohen, 1968) and Krippendorf's alpha is noted as being equivalent to Spearman's *rho* when used in a two-rater ordinal data analyses (Hayes & Krippendorff, 2007). Krippendorf's alpha ranges between 0, indicated complete absence of reliability, and 1 indicating perfect reliability. There is also a weighted version of Gwet's AC_1 (Gwet, 2008) for use with ordinal data, although this is even less commonly used than Krippendorf's alpha.

3.3. Reliability of continuous data

Continuous measurements in animal behaviour will typically take the form of durations or latencies, percentages, or ratings on a visual analogue or Likert scale; these are all commonly treated as continuous data.

Where you have paired data (e.g. from two different raters, the same rater on two occasions or test-retest data from two times), the first port of call for evaluating the reliability of continuous measurements will typically involve the use of exploratory statistics such as scatter plots[1] (Rousson et al., 2002). It would not be suitable to simply calculate a correlation coefficient and consider that to represent agreement, as correlations measure the strength of a relationship and not how well to measures agree (Bland & Altman, 1995). Scatter plots can help the researcher to identify whether there is any systematic bias present (such as one rater scoring consistently lower than the other, or scores improving over time due to habituation or learning) as well as highlight the extent of measurement error (where differences would be randomly distributed either side of the reference line).

If you have identified systematic error in your data, Rousson and colleagues (2002) recommend using different statistics to quantifying reliability depending on how you wish to treat this error. Under a test-retest situation where systematic error is due not to lack of reliability but to learning or development (e.g., most animals scores change in the same direction), reliability can be quantified using a Pearson product-moment correlation, which does not penalize against systematic error. However, for evaluation of inter-rater and intra-rater reliability for continuous data it is recommended to use an intra-class correlation coefficient (ICC), which can take into account systematic bias in measurements that reduce reliability.

3.3.1. Choosing an appropriate ICC

Before you conduct your study, it's important to ensure that you include a large enough sample of subjects, with representative variability in behaviour; a low ICC may indicate low rater agreement, but it can also be caused by lack of variation in the subjects and/or a small sample (Koo & Li, 2016; Lee et al., 2012). There is very little guidance to help the researcher calculate a suitable sample size for a reliability study, however Gwet (Gwet, 2014) provides some useful instructions for determining suitable sample sizes for ICC reliability analyses. First, you need to determine what ICC value you expect to see in your population of raters, once you have this then you can calculate your desired confidence interval width as 0.8 x the expected ICC. Using this approach through a series of data simulations, Gwet shows that for ICC's of above 0.8, acceptable sample size can be as low as 20 measurements (as either 2 raters and n=10, 4 raters and n=5 or 5 raters and n=4). However, for ICC's of 0.7-0.8 you would need 40 measurements (as either 2 raters and n=20, 4 raters and n=10 or 5 raters and n=8), and ICC's of 0.6-0.7 would require 60 measurements (as either 2 raters and n=30, 3 raters and n=20 or 4 raters and n=15).

Intraclass correlation coefficients are derived from the application of analysis of variance (ANOVA) models to data on individuals assumed to be a random sample of a larger population (McGraw & Wong, 1996). The type of ANOVA model you choose for your ICC (see Table 2) will depend on the way in which your data was gathered, which variance you consider relevant and what type of agreement you are looking for. If you have a setup where different pairs/groups of raters score different animals (for example a different pair of zookeepers at each zoo), there is no way to estimate the error attributable to individual raters as it will be tangled up with the animal variance. In this situation, a one-way random effects model is appropriate, and all sources of variance are treated as error (Nichols, 1998).

If, however, you have a group of raters who rate all animals, then rater variance can be estimated as a separate source of systematic variance (Nichols, 1998). Using a two-way ANOVA model, the raters

[1] It's important to ensure that your axis units are shown using the same minimum and maximum values when doing this.

can then be accounted for as a second factor. Whether you use a random effects or a mixed model will then depend on the source of your raters. If they are considered to be a random sample of raters from a larger population (for example a group of assistance dog trainers selected randomly from all assistance dog trainers), then a two-way random effects model would be used, and the results may be considered generalisable to the wider population. If this isn't the case, then a two-way mixed effects model should be used and the results considered applicable only to the raters used in the study (Nichols, 1998).

Table 2. Adapted from (Nichols, 1998) indicating under what situations each type of ICC model should be used.

Model Type	For use when
One-way random effects model	Animals are scored by different pairs/groups of raters (i.e., each pair/group of raters scores a subset of animals).
Two-way random effects model	A set number of raters score all animals, and the raters represent a random selection from a larger population. Inferences may be made about the larger population.
Two-way mixed effects model	A set number of raters score all animals, and the raters do not represent a random selection from a larger population. In this case, inferences are applicable only to the raters in the study.

The final factor to consider in selecting the appropriate two-way ANOVA model is whether you need to estimate absolute agreement, or consistency (for one-way models only absolute agreement can be calculated). If systematic differences between raters is assumed, but not relevant to the judgement of 'agreement', then a consistency model can be used. For example, where rater 1 consistently assigns higher scores than rater 2, but comparative agreement is sought, i.e., the raters vary in the same way, assigning higher or lower scores comparatively, as they score different individuals. The definition of agreement by consistency would consider the following set of paired scores: 4:6, 8:10, 6:8, to be in perfect agreement (ICC=1.0) (McGraw & Wong, 1996). Incidentally, ICC statistics with consistency would be the appropriate method to use if you intend to evaluate test-retest reliability whilst accounting for normative population level change between your tested time points, such as when comparing juvenile to adult scores with the same cohort of animals. If, however, the aim of your comparison is to evaluate whether different raters will assign the exact same score, then absolute agreement would be sought. Absolute agreement is particularly appropriate for use in estimating intra-rater reliability, where the animal being scored can be considered not to have changed, and the measurements for each animal are taken from the same rater on more than one occasion.

3.3.2.Interpreting the output: which ICC statistic do I report?

If your intention is to use your measurement tool for applied purposes, the choice of which coefficient to report must be made based upon the intended future usage of the tool. If, however you only wish to show the reliability of the data for your own study and aren't making recommendations for the future use of your measurement tool, the choice will depend on the setup of raters and study in your study.

Although for inter-rater studies, each animal must be rated by more than one rater, the *single measure* ICC statistic should be reported when the intended use of your tool is for data on individuals to be collected by a single person, i.e., where one animal has one rater. Very often there will be one rater scoring all the data, whilst a subset may be scored by two raters to enable reliability testing, in which case, *single measure* statistics are most appropriate. Alternatively, you may have multiple raters contributing data to the main study, but if the individual animals are each rated by a single person, then the *single measure* is the statistic you need to report. Single rater situations are often the case in animal behaviour research where resources are constrained limiting the number of raters that can be used, or research sites (and thus raters) are separated by distance or time meaning different people are needed to score all the subjects. The *single measures* statistics should also be reported for intra-rater reliability.

When you intend to have each individual animal scored by a set of raters and you wish to take an average of their scores, the *average measure* ICC statistic should be reported. Creating aggregate scores in this way helps to minimise the error introduced by the individual rater, but it is resource heavy due to the need for groups of trained raters so is often not feasible for implementation in practice. If you're using an ICC to calculate the average level of temporal consistency over multiple measures, such as for example three measurements between infancy and adulthood, then then the *average measure* statistics would be suitable to report.

Unlike Kappa statistics, there are no accepted rules of thumb for interpreting ICC statistics. When reporting an reliability analysis using an ICC, it's imperative that the researcher report the software used, the model used (as per Table 1), the definition (consistency or absolute agreement) and the type of coefficient (average or single measure) (Koo & Li, 2016; Lee et al., 2012). The reason these must be reported is because the model used and type of coefficient will impact the magnitude of the ICC; average measure coefficients, consistency definitions, and two-way fixed models for example will all yield higher ICC values than their counterparts (Lee et al., 2012).

For those of you interested in how the model you choose impacts how you can infer from the ICC statistics beyond this guide, McGraw and Wong (1996) provide an in-depth discussion of forming inferences from ICC's. Further reading on how to select and report ICC's and can found in (Koo & Li, 2016).

3.3.3.The Bland-Altman Plot

Quantifications of overall reliability based upon correlations may not take into account the degree of difference between raters in terms of true agreement. If this is important to you, a plot that has utility for evaluating inter-rater reliability is the Bland-Altman 95% limits of agreement plot. Originally designed for evaluating agreement between two different measurement methods in medical sciences (Bland & Altman, 1995; J.M.Bland & D.G.Altman, 1986) the Bland-Altman plot can be useful for interpreting the magnitude of agreement between specific pairs of raters. Here, the researcher calculates the difference in scores between two raters, and the mean of the two raters scores, then plots the difference against the mean. The mean difference and standard deviation of the differences are then indicated on the plot to provide the 95% limits of agreement. The data plotted in Fig. 3 has an almost perfect correlation 0.99, but you can see from the Bland-Altman plot that the two sets of data are not in perfect agreement. The mean difference for the two fake ratings is -0.36 and there is a critical difference of 1.5 units either side of the mean difference, within which 50% of ratings will fall. Using this method for inter-rater reliability allows you to see how many pairs of raters agree well (i.e., with a difference close to zero), or not at all, and to what extent. As with any measure of agreement, it is important for the researcher to decide *a priori* what would be an acceptable width for an agreement interval based upon biological or clinical utility. This method assumes independence of the difference from the mean difference, and normality in data distribution, but steps can be taken to transform the data such as logarithmic transformations (for more information about the Bland-Altman method see references (Bland & Altman, 1995; Giavarina, 2015)).

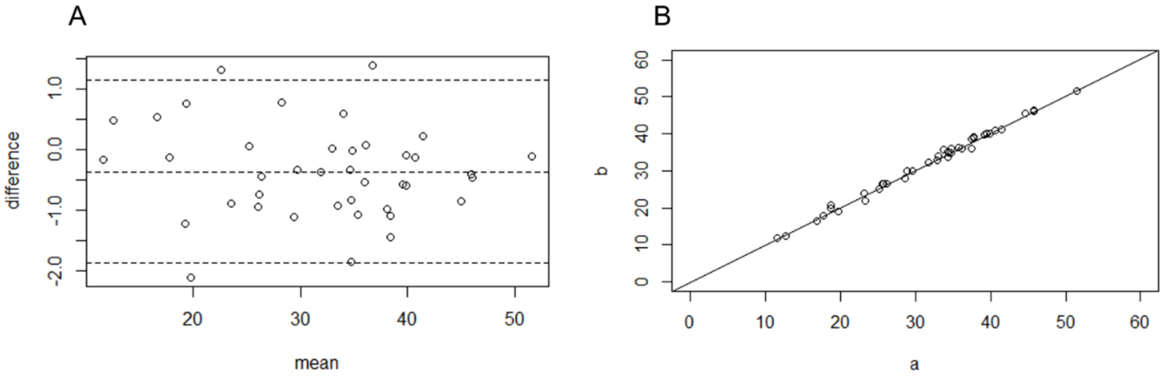

Figure 3. Example of a Bland-Altman plot (A) versus a scatter plot (B) using the same hypothetical data generated in R via the BlandAltmanLeh package (Lehnert, 2015). The dotted line in the middle of the Bland-Altman plot indicates the mean difference of -0.36 and the dotted lines above and below indicate the 95% limits of agreement.

4. TEST-RETEST RELIABILITY FOR PERSONALITY ASSESSMENT

In the case of test-retest reliability, you may expect to see systematic differences in your population's measurements between the first and second test. Such differences may be expected due to learning or habituation from repeated exposure, or they may be developmental if you are evaluating an ageing population. As mentioned previously, a simple Pearson correlation can be used as a reliability coefficient in the case of normally distributed continuous data with systematic bias (Rousson et al., 2002). However, when assessing test-retest reliability for the purpose of personality assessment there are other considerations.

Systematic bias such as that seen in an ageing population doesn't mean that individual consistency (part of the definition of personality) isn't still present. One such phenomena is 'normative change', which is where individuals are expected to change in a similar way in response to development or ageing (Mccrae et al., 2000; Wallis et al., 2014). In such cases, a test for differences, such as a t-test, would identify population-level shifts in behaviour, such as for example a reduction in play behaviour with increasing age. Significant mean differences between tests do not mean that the animals' behaviour is not consistent; consistency may still be evident if individuals within the population have maintained their rank order position in relation to each other. If, for example, we assume play behaviour to be consistent between individuals, and animal A was more playful than animal B as a juvenile, then animal A should still be more playful than animal B as an adult; despite all animals in the population exhibiting less play behaviour between the juvenile and adult stages in general. In such a case, a non-parametric correlation using Spearman's rank (or Kendall's Tau-B if you expect to have ties such as with an ordinal scale) could be used to estimate the magnitude of inter-individual consistency present over time in your test-retest sample.

5. PRACTICAL CONSIDERATIONS FOR RELIABILITY TESTING

As has been mentioned previously in this guide, there is no definitive threshold for acceptable reliability. The decision about what to accept should be made prior to data analysis and should be tailored the specific context of your analysis.

In some cases, it may be expected that agreement would be weak (e.g., single measure ICC's of <0.40 or average measures of <0.50) such as when different measurement methods are being used. Another scenario where low inter-rater agreement may be expected, specific to captive animal behaviour assessment, could be when your raters are people with differing experience of, or relationships to, the animal in question. An example of this from my own work was a study that evaluated agreement between lay people who lived with trainee guide dogs and dog behaviour trainers who only saw the animals for an hour a week, using two similar but slightly different dog behaviour questionnaires (Harvey, Craigon, Blythe, et al., 2016). Not only were the methods of assessment different, but the rater's knowledge of dog behaviour was different, as was their relationship to the animals, so it could be reasonably expected that the dogs may behave differently for the different raters (Horn et al., 2013; Kerepesi et al., 2015). The various factors influencing measurements in this situation would likely reduce agreement considerably and led us to set an acceptable threshold of >0.30 significant at 95% confidence from single measure ICC (using a 2-way mixed consistency model). In this case no averages were ever going to be made from combined scores, so the purpose of the comparison was to show that there was some degree of significant overlap in the behaviours observed by distinctly different people.

For test-retest reliability of behavioural data, the threshold for acceptance will depend on the parameters of the study. Shorter time periods between testing or observation will be expected to produce larger estimates of reliability, although diurnal patterns may introduce biases if not controlled for. Reliability coefficients will also be expected to be lower if your study subjects are juvenile or not behaviourally mature due to on-going neurological and cognitive development (see (Harvey, 2019) for an overview of the factors impacting animal behaviour during adolescence). In a meta-analysis of temporal consistency of personality measures in dogs, Fratkin and colleagues (2013) showed personality measures were more consistent in older dogs, when intervals between assessment were shorter and when the same measurement tool was used on both occasions.

Another consideration is whether the purpose of your study testing both inter- and intra-rater reliability. For studies where just one person does all behaviour scoring, the data being used is limited to the study in question, and there is no suggestion that others will use the same methods in the future for applied purposes (such as an analysis within a behavioural ecology PhD project) it would be suitable to test for intra-rater reliability, with no need to test for inter-rater reliability. However, for any study where behavioural data is collected by more than one rater, or if there is an intended applied use of the behaviour collection protocol by future persons, inter-rater reliability testing will be imperative.

6. ACKNOWLEDGEMENTS

Thank you to Dr Giuliana Miguel-Pacheco, Dr Rachel Casey and Dr Melissa Upjohn for reviewing this and convincing me it was suitable to publish.

7. References

Bland, J. M., & Altman, D. G. (1995). Comparing methods of measurement: why plotting difference against standard method is misleading. *The Lancet*, *346*, 1085–1087. https://doi.org/10.1016/S0140-6736(95)91748-9

Charman, C. R., Venn, A. J., & Williams, H. C. (2004). *The Patient-Oriented Eczema Measure*. *140*, 1513–1520.

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220. https://doi.org/10.1037/h0026256

Devine, O. (2003). The impact of ignoring measurement error when estimating sample size for epidemiologic studies. *Evaluation & the Health Professions*, *26*(3), 315–339. https://doi.org/10.1177/0163278703255232

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 387–382. https://doi.org/10.1037/h0031619

Fratkin, J. L., Sinn, D. L., Patall, E. A., & Gosling, S. D. (2013). Personality Consistency in Dogs: A Meta-Analysis. *PLoS ONE*, *8*(1). https://doi.org/10.1371/journal.pone.0054907

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, *25*(2), 141–151. https://doi.org/10.11613/BM.2015.015

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*, 29–48. https://doi.org/10.1348/000711006X126600

Gwet, K. L. (2014). Handbook of Inter-Rater Reliability: the definitive guide to measuring the extent of agreement among raters. In *Gaithersburg, MD: STATAXIS Publishing Company* (4th Ed). Advanced Analytics, LLC.

Harvey, N. D. (2019). Adolescence. In *Encyclopedia of Animal Cognition and Behavior* (pp. 1–7). Springer International Publishing. https://doi.org/10.1007/978-3-319-47829-6_532-1

Harvey, N. D., Craigon, P. J., Blythe, S. A., England, G. C. W., & Asher, L. (2016). Social rearing environment influences dog behavioral development. *Journal of Veterinary Behavior: Clinical Applications and Research*, *16*, 13–21. https://doi.org/10.1016/j.jveb.2016.03.004

Harvey, N. D., Craigon, P. J., Sommerville, R., Mcmillan, C., Green, M., England, G. C. W., & Asher, L. (2016). Test-retest reliability and predictive validity of a juvenile guide dog behavior test. *Journal of Veterinary Behavior: Clinical Applications and Research*, *11*, 65–76. https://doi.org/10.1016/j.jveb.2015.09.005

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, *1*, 77–89. https://doi.org/10.1080/19312450709336664

Horn, L., Range, F., & Huber, L. (2013). Dogs' attention towards humans depends on their relationship, not only on social familiarity. *Animal Cognition*, *16*(3), 435–443. https://doi.org/10.1007/s10071-012-0584-9

J.M.Bland, & D.G.Altman. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*, 307–310.

Kerepesi, A., Dóka, A., & Miklósi, Á. (2015). Dogs and their human companions: The effect of familiarity on dog-human interactions. *Behavioural Processes*, *110*, 27–36. https://doi.org/10.1016/j.beproc.2014.02.005

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

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

Lehnert, B. (2015). *Plots (Slightly Extended) Bland-Altman Plots* (0.3.1). CRAN. https://cran.r-project.org/web/packages/BlandAltmanLeh/BlandAltmanLeh.pdf

Mccrae, R. R., Costa, P. T., Hrebickova, M., Avia, M. D., Sanz, J., Sanchez-bernardos, M. L., Kusdil, M. E., Woodfield, R., Saunders, P. R., & Smith, P. B. (2000). Nature Over Nurture : Temperament , Personality , and Life Span Development. *Journal of Personality and Social Psychology*, *78*(1), 173–186.

McGraw, K. O., & Wong, S. P. (1996). Forming Inferences about Some Intraclass Correlation Coefficients. *Psychological Methods*, *1*, 30–46. https://doi.org/10.1037/1082-989X.1.1.30

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/bm.2012.031

Nichols, D. P. (1998). *Choosing an intraclass correlation coefficient. SPSS Inc.* http://www.ats.ucla.edu/stat/Spss/library/whichicc.htm

Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater and test-retest reliability of continuous measurements. *Statistics in Medicine*, *21*(22), 3431–3446. https://doi.org/10.1002/sim.1253

Ruefenacht, S., Gebhardt-Henrich, S., Miyake, T., & Gaillard, C. (2002). A behaviour test on German Shepherd dogs: Heritability of seven different traits. *Applied Animal Behaviour Science*, *79*(2), 113–132. https://doi.org/10.1016/S0168-1591(02)00134-X

Svartberg, K., & Forkman, B. (2002). Personality traits in the domestic dog (Canis familiaris). *Applied Animal Behaviour Science*, *79*(2), 133–155. https://doi.org/10.1016/S0168-1591(02)00121-1

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*(5), 360–363.

Wallis, L. J., Range, F., Müller, C. A., Serisier, S., Huber, L., & Virányi, Z. (2014). Lifespan development of attentiveness in domestic dogs: Drawing parallels with humans. *Frontiers in Psychology*, *5*(FEB). https://doi.org/10.3389/fpsyg.2014.00071