

## **BRAVO - a Workflow for Improving Rating Reliability in Behavioural Research**

Damien Neadle<sup>1,2</sup>, Alba Motes-Rodrigo<sup>3</sup>, Sarah R. Beck<sup>2</sup>, and Claudio Tennie<sup>4,5</sup>

<sup>1</sup> Birmingham City University, Department of Psychology, Birmingham, UK

<sup>2</sup> University of Birmingham, School of Psychology, Birmingham, UK

<sup>3</sup> University of Lausanne, Department of Ecology and Evolution, Lausanne, Switzerland

<sup>4</sup> University of Tübingen, WG Early Prehistory and Quaternary Ecology, Tübingen, Germany


<sup>5</sup> Words, Bones, Genes, and Tools: DFG Center for Advanced Studies, Tübingen, Germany

---


### Abstract

Reliability assessments are a quality control protocol commonly employed in fields of research that deal with video-recorded behavioural data. During these assessments, the same sample of videos is coded (at least) twice by the same researcher (intrater reliability), or - more often - by two different researchers independently (interrater reliability). Next, levels of agreement are quantified to assess how reliable the behavioural classification is. In this manuscript, we concentrate on interrater reliability, though our points hold generally true for both cases. Despite the importance of interrater reliability assessments to ensure research quality, to the best of our knowledge there is no guideline to date specifying how they should be conducted to avoid potentially detrimental effects of ‘coders’ degrees of freedom’ (CDF) and ‘questionable coder practices’ (QCP). For instance, there is no consensus regarding how large the sample of behaviours evaluated should be, the sample composition, the inclusion of negative controls or what statistical measures should be used to

---

Damien Neadle  0000-0001-8559-436X

Alba Motes-Rodrigo  0000-0002-4444-7723

Sarah Beck  0000-0001-6426-1603

Claudio Tennie  0000-0002-5302-4925

The Authors declare no conflicts of interest. During the process of designing this methodology, Damien Neadle received funding from the Economic and Social Research Council, under a full PhD Studentship ES/J50001X/1. Claudio Tennie was supported by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63) and also received funding from the European Research Council under the European Union’s Horizon 2020 Programme (H2020-EU.1.1.) / ERC grant agreement No. 714658. During the writing stage, Claudio Tennie also received support from the DFG-Centre for Advanced Studies in the Humanities Words, Bones, Genes, Tools (DFG-KFG 2237). We wish to thank two anonymous reviewers for very helpful and constructive criticisms.

- ✉ Claudio Tennie, Eberhard Karls Universität Tübingen, Institut für Ur- und Frühgeschichte und Archäologie des Mittelalters, Abteilung für Ältere Urgeschichte und Quartärökologie, Hölderlinstrasse 12, 72074 Tübingen, Germany. E-mail: [claudio.tennie@uni-tuebingen.de](mailto:claudio.tennie@uni-tuebingen.de)

compare the raters' classifications. To begin to fill this methodological gap, we provide a list of best practices to conduct reliability tests, which we term the BRAVO (Balanced Reliability Assessment of Video Observations) workflow. We complement these recommendations with a series of simulations highlighting the properties of BRAVO and its use-cases. BRAVO represents the first step in creating a methodological gold-standard that researchers can use to perform valid reliability assessments. Given the widespread use of behavioural data across fields, we hope that the BRAVO workflow will be implemented by researchers from a variety of disciplines such as psychology, ethology, behavioural economics, and anthropology to increase quality control and scientific transparency.

*Keywords:* reliability, replication crisis, classifier validity, simulation, behavioural sciences

---

## Introduction

Across fields of research, scientists must ensure that their results undergo quality controls with the aim that human errors or biases are not influencing their findings. One such control involves ensuring that the metrics employed to quantify effects are valid and that they capture what they are intended to measure. Furthermore, validity must not be transient and should always be paired with reliability, meaning that not only metrics but the measurements themselves need to be as objective as possible. Reliability is the consistency of a measure across its use, whether within individual raters across time (intrarater) or between different raters (interrater). Although the latter is often the focus in quality checks performed by psychologists, both are important. Interrater reliability (IRR) reflects the degree to which rater X and rater Y agree on what they observed and is often calculated using the Kappa statistic (McHugh, 2012).

Despite their frequent use across scientific fields, the processes of acquiring and preparing the samples used to calculate IRR often varies from discipline to discipline and, perhaps of more concern, between research groups. There are, in other words, 'coders degrees of freedom' (CDF) at work. Coders' degrees of freedom (note, the term 'coder' here can be considered analogous to 'rater' which we use as per the accepted inter/intra-rater reliability) may reduce the validity and generalisability of reliability measures as well as lead (intentionally or unintentionally) to 'questionable coder practices' (or QCP). This situation mirrors researcher 'degrees of freedom' and 'questionable research practices' (e.g., John et al. 2012), which strongly contribute to the so-called replicability crisis in psychology and beyond. In this paper, we present a list of recommendations intended as a first step towards the standardisation of IRR calculations based on observations in behavioural research. Specifically, we focus on how to prepare the samples for IRR assessments in order to curb the associated coders degrees of freedom at this stage of the process.

Observational data refers to behavioural annotations of a phenomenon/ phenomena as perceived by a researcher. In our article, most examples of observational data will involve comparative psychological studies (generally of non-human primates and children due to our own research foci), but the proposed guidelines and conclusions are generalisable across fields that use behavioural data from video.

While issues of metric validity and reliability across observers are not exclusive to psychology and ethology, to our knowledge and following a targeted literature review, “best practices” for ensuring consistency of scientists’ measurements and observations are similarly lacking across fields. For instance, although IRR on doctors’ diagnosis is often performed (Mohan et al., 2017), we were unable to find published guidelines on how to obtain the data used to perform IRR or how to perform unbiased IRR in healthcare research. In positivist domains, however, there have been significant attempts to standardise practices dealing with qualitative data. Here we refer to the extensive literature surrounding Thematic Analysis, and the objectification of this methodology using either coding reliability or ‘codebook’ approaches (Byrne, 2022; Roberts et al., 2019). Nevertheless, these approaches are not without critique. Key proponents of Thematic Analysis remain sceptical about the validity of the approach, as any attempt to remove researcher degrees of freedom in interpretivist spheres may remove the inherent reflexivity of the method and its associated benefits (Braun & Clarke, 2022). This said, having standardised methodologies for the ‘codebook’ approaches (discussed by Byrne, 2022) does allow a foothold for researchers and gives a clear point of reference for practitioners who wish to adopt this theoretical stance to align with others in the field. This goal is similar to that of the workflow described below.

In the following sections, we describe the different metrics employed to quantify reliability, the different methods used to generate the samples where these metrics are applied, their respective advantages and disadvantages and the targets of IRR. We then describe the BRAVO (Balanced Reliability Assessment of Video Observations) workflow and present a series of simulations showcasing its uses and characteristics.

## **Reliability Metrics**

As noted above, reliability in research is an umbrella term and it can be used to describe different processes. Reliability can refer to the consistency of a psychometric measure or other such questionnaire across test items. This type of reliability is often calculated using the Cronbach’s Alpha (e.g., Tavakol & Dennick 2011), which is a metric ranging from 0 to 1 quantifying how much test items are correlated or covary. Given that the calculation of Cronbach’s Alpha is largely automated, this metric is relatively free from bias during calculation.

*Intrarater* reliability, refers to the degree to which rater X’s decisions about what is/is not behaviour A changes throughout their coding process, e.g., due to sampling variance, noise, chance, learning, boredom or distractions. *Intrarater* reliability often relies on a predefined list of behavioural categories and descriptions/photographs, called ethograms, which must be used consistently during the coding process.

*Interrater* reliability (IRR) refers to the level of agreement between independent observers regarding what constitutes a behavioural instance, when it takes place and/or how long it lasts. In recent years, several articles have been produced that support users in their decision of an appropriate statistic (see Harvey, 2021 for a helpful discussion). Usually, IRR is calculated using the Cohen Kappa’s statistic (Cohen, 1968) which is calculated as demonstrated below (Table 1; where in the example  $\kappa = .72$  and  $SE(\kappa) = 0.007$ ).

**Table 1**

*Coding Agreement Matrix for Illustrative Example. Note That Figures Here Are Arbitrary*

		R2		
		Yes	No	
R1	Yes	48	2	50
	No	12	38	50
		60	40	

$$p_e = (R1_{yes} \times R2_{yes}) + (R1_{no} \times R2_{no})$$

$$p_e = \left(\frac{50}{100} \times \frac{60}{100}\right) + \left(\frac{50}{100} \times \frac{40}{100}\right)$$

$$p_e = 0.3 + 0.2$$

$$p_e = 0.5$$

$$p_o = \frac{n_{agreements}}{n_{observations}}$$

$$p_o = \frac{(48 + 38)}{100}$$

$$p_o = 0.86$$

$$k = \frac{p_o - p_e}{1 - p_e}$$

$$k = \frac{0.86 - 0.5}{1 - 0.5}$$

$$k = 0.72$$

$$SE(k) = \sqrt{\frac{p_o(1 - p_o)}{n(1 - p_e)^2}}$$

$$SE(k) = \sqrt{\frac{0.86(1 - 0.86)}{100(1 - 0.5)^2}}$$

$$SE(k) = 0.0069$$

Initially, Cohen (1960) designed the calculations of the Kappa statistic around just two raters. If more raters are used, *multiple* pairwise calculations between all raters would be needed (Sainani, 2017; i.e., A-B, A-C, B-C), thus increasing the familywise error rate over a sample A-B. An alternative that avoids the issue of multiple testing is to use an extension of Cohen's Kappa (e.g., Fleiss's Kappa) or the Intraclass Correlation Coefficient (ICC; Hallgren, 2012). ICC differs from Kappa and Fleiss' Kappa in that it handles continuous data and multiple raters, measuring both consistency and absolute agreement among them. Furthermore, ICC accounts for the magnitude of disagreements among raters, while Kappa methods are primarily for categorical data and absolute agreement. Yet, overall, there seems little to no 'standard' (let alone *ideal*) number of raters required for IRR as a whole (see a review by Barth et al., 2016 from the field of disability research where the authors report a range of 2–106 raters across 23 studies; *Mdn* = 12). However, given the relative ubiquity of Cohen's Kappa in behavioural studies (Konstantinidis et al., 2022) it would appear as if the *current de facto* field standard is set at two raters. Consequently, we shall assume two raters for the rest of this article. This said, it should be noted that when resources allow, three or more reliability raters that are naïve to the goals of the study should be used. As a reviewer of this article helpfully suggested, if more than two raters are available, their scientific backgrounds should differ in order to ensure that coding performance is not dependent of specific training. Otherwise, the coding risks to run into some sort of 'pseudoreliability' problems (e.g., imagine raters agree mainly due to matching backgrounds, such as their common training in a single lab).

The point of second (and beyond) rater naivety is one worth expanding on, as it can have a significant impact on the results of the study (Hróbjartsson et al., 2012). In many cases, second coding for IRR assessments will likely be carried out by other members of a lab (often a student or research assistant, and generally akin to convenience sampling). However, by simply taking those people in close proximity (and, often, matching background) to the first rater, it is possible that the reliability of the second rater's data could be inflated. For instance, consider that a common practice is to have regular 'lab meetings' where members discuss ongoing projects. Imagine the primary researcher (and, likely, first rater) were to discuss new literature that suggests wild bonobos (*Pan paniscus*) can measure liquid quantities (to use an imaginary example), and that based on this research they have designed a study to test captive bonobos for this ability. The second rater may remember this information and thus may be *more likely* to code certain liquid interactions in the captive sample as liquid measurements. These biases may manifest in a number of ways, including (but not limited to): over-interpretation of ambiguous data, selective attention to certain aspects of the data or unintentional emphasis on data that aligns with the alternative hypotheses. It is for these reasons that it is important that the decision of *who* should act as reliability raters is not taken lightly, and consideration is given to prior interactions that might bias the coding.

In terms of the number of observations to code, Konstantinidis et al. (2022) provide an interesting simulation that provides further support to prior work (Wongpakaran et al., 2013) suggesting that in smaller or unbalanced samples, Cohen's original statistic may be unsuitable, e.g., sizes lower than 5 observations or with substantial variance in the numbers of observations, as in these cases the variance increases. This proved to be the case for many of the IRR methods simulated by Konstantinidis et al. (2022). In cases where very small sample sizes are unavoidable (e.g., for rare behaviours), Gwet's AC1 should be applied, which is designed to avoid the two cases described, the so-called "prevalence" and "bias" problems (Di Eugenio & Glass, 2004).

As will become clear throughout this paper, the currently employed methods of assessing IRR are far from consistent in various fields, even to the extent that there is no standard protocol to report *how* the Kappa value was reached. This means that the protocols for IRR cannot often be readily compared. In the next section, we describe some of the commonly used methodologies to generate the behavioural samples that raters use to classify behaviours and apply the abovementioned metrics.

### **Video Sample Generation Methods**

One method used to generate behavioural samples to perform IRR assessments is what we term here the 'clipping method'. This method involves the first (main) rater coding the videos according to the predefined (or continually evolving) ethogram and noting all relevant behavioural instances and contextual information (e.g., timestamp, handedness, social partners, etc.) as they occur. Once a behavioural dataset has been fully built by, usually, this main rater, a subset of these behaviours (usually about 5–20% of instances, although no standardised protocol exists, to the best of our knowledge), are given to a second (reliability) rater who then uses the same (or "final") ethogram to code the provided behavioural sample. The data generated from the subset of behaviours by the second rater is then compared to the data generated by the main rater on the same subset, using the Kappa statistic (Cohen, 1968) to assess IRR.

A clear advantage of the clipping method is that it allows researchers to filter out 'noise' in the data before coding, meaning that behaviours outside the scope of the study can be ignored (at the discretion of the main rater). In this context, clipping is the act of taking a sample of a video clip of a behaviour from a longer video recording of a trial. This has the distinct advantage that second raters need not be experts in the target species' entire behavioural repertoires to identify behaviours of interest. This adds a level of granularity to the process and therefore – everything else being equal – should improve reliability. However, the clipping method has the disadvantage that the researcher can bias the second rater, either by selecting only clear-cut behavioural examples, ignoring context or, most concerningly, guaranteeing that every clip will contain *some* target behaviour. This last point is

perhaps the Achilles heel of an otherwise elegant method, where the second rater may be biased by knowing or guessing that all clips contain a behaviour of interest. We discuss this issue further below.

An alternative to the clipping method, which we term the ‘time selection method’, removes this specific limitation, by basing the selection of the behavioural subset provided to the second rater not on content, but on time. For example, in a study with 100 hours of video, a subsample of, say, 20 hours (20%) may be provided to the second rater. This subset can be randomly or systematically selected, e.g., 40, randomly selected, 30-minute clips can be provided or the (randomly selected) first/middle/last  $n$  minutes of a recording can be used. Here, the advantage is that there is no possibility of a forced and informed choice, leading to a potential type-one error, if the clips contain ‘no behaviour’ or ‘irrelevant behaviours’ as much (if not more) as they do target behaviours (which may be often, though not always, the case). The use of ‘dummy clips’ is a particular advantage in exploratory research studies, where the main rater might miss behavioural occurrences, due for instance to coding fatigue, bias or lack of a search image (see intra-rater reliability). However, the time selection method has the disadvantage that a second rater might fixate on irrelevant behaviours – especially considering that the second rater should be naïve to the study goals. For example, a study of object manipulation might have a naïve second rater focus on feeding behaviour, mistaking it for object manipulation (e.g., “object manipulation of vegetation”). Equally, this method has the very real risk that behaviours integral to the study may be omitted from IRR assessments entirely owing to random video selections. Finally, the biggest disadvantage of this method is what we term the ‘granularity problem’ (similar to the phenomenon in cultural evolution described by Cartmill & Byrne, 2011; Charbonneau & Bourrat, 2021).

Here, the ‘granularity problem’ is the issue that for the time selection method to be unbiased, both raters need to provide timestamps of the behaviours that they observe. This might seem a trivial problem at first; however, the reliability of the study then depends, at least in part, on the degree to which the ethogram specifies when a behaviour starts and/or ends. For example, one might specify that behaviour X is coded when an individual touches object A, but this relies on the degree to which one can be accurate in the time when the individual first touches the object and also what qualifies as a ‘touch’: Is it brushing it with the hand, taking a two-second break to eat, then beginning behaviour X? Or does it begin at the second contact with object A? These intricacies *can* be ironed out, theoretically, with a very well-defined ethogram, but it may involve several iterations of back-and-forth communication with the second rater to fully understand these potential issues. This need for detailed explanations to the second rater is a problem in itself as it can introduce additional problems associated with overly long ethograms as well as (potentially undocumented) communication-led agreement inflation. Consequently, due to the granularity problem, the second rater might set the starting time of a behaviour earlier/later than the main rater, which in our example of a 30-minute clip, can build

up to low agreement and low Kappa values (e.g.,  $\kappa = .20$ ; “acceptable” Kappa is generally  $\kappa > .60$ ; Cohen, 1968; McHugh, 2012).

In cases where the goal of the study is to evaluate the timing (e.g., latency) or length of a behaviour, the clipping method would often not suffice. In this instance, a challenge that researchers might face is to define a degree of tolerance in the timestamp data, which does not undermine the validity of the data without sacrificing the validity of the reliability. Similarly, it is important to consider the fact that in some cases, these subset preparation methods would not work, such as in animal field studies, where it may sometimes not be possible to collect video footage in sufficient quality/quantity. In this and other similar cases where videos are scarce, it is perhaps important to consider multiple simultaneous observers (e.g., Perry, 1995; Perry et al., 2008), who can then afterwards compare live behavioural classifications through Kappa or have an ethogram that has been subjected to peer review/feedback via a registered reports/pre-registration process. This said, these issues are beyond the scope of this article and should be treated as mere suggestions.

### **Reliability Targets and the BRAVO Workflow**

An important, and often-overlooked, fact is that IRR *must* be gained for all relevant aspects of a dataset. That is, if a study on object manipulation analyses data on behaviour type, handedness, object type, behavioural duration and the social context of the behaviour, it is not sufficient to provide a single Kappa value. Each variable included in the analyses needs to be checked for IRR and acceptable Kappa values must be achieved before conducting statistical analyses. Similarly, the methods used to assess IRR need to be fully and clearly reported. To guide and help researchers in how to conduct and report these IRR assessments for different types of data, we have created the BRAVO workflow (Balanced Reliability Assessment of Video Observations; see Figure 1). To illustrate how to apply this workflow and the justification for its recommendations, we have further conducted a series of simulations using artificial data (see S2<sup>1</sup> for script) based on the observational study of *Pan* object manipulation by Koops et al. (2015). In this study, subjects were observed for a set period of time and their behaviours were coded against a predefined ethogram. For the purpose of our own paper, we borrowed the fictitious species of ‘oranzees’ and their behavioural repertoire from Acerbi et al. (2022; see ethogram, Table 2 and S1<sup>1</sup> for descriptions of behaviours). The reasons for using simulations are first, to illustrate how the BRAVO workflow can be applied to realistic datasets (of any size or complexity) and second, to justify the recommended sample sizes required to conduct IRR which scientists can refer to in future publications. We hope to bring to attention the potential problems of coders degrees of freedom and questionable coder practices (QCP) as well as contribute to the

---

<sup>1</sup> Supplements are available in online supplementary material.



further improvements of reliability analyses in ours and other research fields that employ behavioural data.

In our simulations we assume the current standard in the field, i.e., that there is a single main rater (who can, for now, still be non-naïve to the study). We do however suggest that it would be a good idea to determine whether intrarater reliability is sufficient in this main rater. Thus, the main rater should code a sample of data at the start of the coding process and then code that same data at the end to determine whether their coding changed throughout the process. For now, we regard this as optional, however, because our goal here is to initiate a process of standardization of IRR rather than a gold-standard procedure. Such a procedure (or set of procedures) is not only outside the scope of our current manuscript but may (at present) be too demanding on research teams. Instead, we opted to tackle one of (what appears to us) most important problems in IRR assessments: the preparation of samples.

## **Simulation Design**

We conducted a simulation to generate realistic datasets for demonstrating the application of the BRAVO workflow to IRR assessments. We simulated the behaviour of 20 ‘oranzees’, labelled A through T, which were observed individually for one 10-minute focal period (600 seconds). During this period, a subject could be observed performing none, one, or multiple behaviour(s) from a predetermined ethogram at random times (see below) in what is called a focal follow (Altmann, 1974). If the 10-minute focal period ended and a behaviour was still being performed, then the trial continued until the behaviour had finished. The behaviours of the focal subject could be social or food-related (the simulation drew from a pool of each behavioural sub-category), and within these two broad categories the behaviours could be further classified in four sub-categories each (see Table 2; the attribution of behaviour type was based on a random draw from this ethogram). In addition, each of these sub-categories contained various behaviours adding to a total of 40 different variants (e.g., “Air-Split” and “Tongue-Bathe”; see Table 2 for full list; again, these were random drawn). Behaviours were also coded as left or right-handed and in a group or individual (50% chance each). The duration of the behaviour was a random draw from a Gaussian distribution bound between 1 and 600 seconds.

For the purpose of this simulation, there is an inherent assumption that the ‘first rater’ was ‘correct’ in their classification of the behaviour. This is because there are no ‘true’ clips to code from. This is likely a somewhat utopian assumption as even the best raters commit errors. However, we accept this assumption because in real IRR assessments, if good reliability is achieved after comparing the second and the first raters’ classifications, typically the first rater’s data is analysed in the paper. This means that even in real studies it is assumed that the first rater’s classification is correct.

**Table 2**

*Behavioural Ethogram of Behavioural Categories, Sub-Categories, and Individual Behaviours From the Oranzee Repertoire (All Evocative Terms Only)*

Behavioural category	Behavioural sub-category	Behaviour
Social	Play	Fruit-Missile
		Slap-Fight
		Air-Split
		Leaf-Mask
		Whistle
	Display	Pebble-Tease
		Stone Drop
		Branch Pull-Release
		Arm-Cross
		Two-Hand-Drum
	Groom	Splash
		Arm-Swing
Tool Back-Scratcher		
Hand Back-Scratcher		
Tongue-Bathe		
Courtship	Tooth-Pick	
	Dirt-Shower	
	Ant-Shower	
	Flower-Offer	
	Hand-Stand	
Food-related	Fruit-Hammer Foraging	Rope-Swing
		Leaf-Fan
		Wreath-Clutch
		Ear-Pull
	Stick-Based Foraging	Wood-Wood
		Wood-Stone
		Stone-Wood
	Anvil Smash	Stone-Stone
		Stick-Throw V
		Stick-Throw A
Rolling Pin Techniques	Fish-Stab	
	Hedgehog-Flick	
	Anvil-Smash S	
		Anvil-Smash W
		Smash-Ground
		Drop-Ground
		Rolling-Wood
		Rolling-Stone
		Rolling-Bone
		Rolling-Other

*Note.* Adapted from Acerbi et al. (2022) for ease of simulation

Owing to the nature of the simulation, it was possible for subjects to engage in multiple behaviours within the same ‘bout’; e.g., oranzee C could engage in tool back-scratcher between 429 and 455 seconds, and ant-shower between 432 and 462. This is because the start times of each behaviour were integers randomly drawn from

a Gaussian distribution limited between 1 and 600 seconds; meanwhile the durations were randomly drawn from a Poisson distribution with an average duration of 25 seconds. Whilst this feature of the data is a result of the method of our specific simulation, it *does* reflect in many ways typical great-ape observational research in which bouts can consist of multiple behaviours. Importantly, this choice does not compromise the goal of the simulations of generating datasets for IRR.

The final dataset of simulated oranzee behavioural data was structured as a table with 8 columns and 700 rows. The simulation was set similarly to a stratified sampling method. Such methodologies are often employed in IRR to ensure that coding is consistent across behavioural categories. In this case, the decision was made to stratify at the sub-category level, meaning that each individual displayed a behaviour, from each sub-category, six times. However, the behaviours varied between individuals.

**Table 3**

*File Format for the Simulated Data Set*

---

Column	Description
oranzee_id	A-T 35 rows per oranzee
category	Social/food-Related 20 social/15 food-related per individual.
sub_category	Each 'subject' displayed a behaviour from each sub-category 5-6 times.
behaviour	Behaviours were indexed within their sub-category and a random draw determined which behaviour was expressed.
time_start_sec	Random draw from between 1 and 600 to determine start time
duration_sec	Random draw from a Poisson distribution with rate param (average) set to 25
handedness	Random draw between right and left
sociality	Random draw between yes and no

---

### **Size Selection of Sample to be Provided to Second Rater**

It is common practice in the fields of comparative psychology and behavioural ecology that only a subset of behaviours or clips are passed to the second (or further) rater(s) (see van Allritz et al., 2021; Leeuwen et al., 2023; Tecwyn et al., 2023). However, in the medical or psychiatric fields (amongst many others) it is more commonplace to see total (i.e., 100%), 'dual' coding (e.g., Bakar et al., 2017; Denis et al., 2016). Whilst the 'gold-standard' here would be to perform reliability analyses for all behaviours observed, we (and the field as a whole appear to) acknowledge the need for sampling in the interest of practicality (coding requires resources). Again, in these cases it is imperative that a representative and statistically valid sample is selected as per the BRAVO workflow.

According to McHugh (2012), Kappa comparisons should not be performed with fewer than 30 data points from each rater (owing to the likelihood of very large

confidence intervals [CI] and therefore false negatives). Thus, if using the clipping method (see above), where each clip contains a single behaviour, at least 30 clips would be needed. This said, McHugh (2012) also suggests that a sample size of more than 1,000 observations is required to have the most mathematically reliable estimate of agreement using the Kappa statistic. Given the breadth of the range this creates, we calculated the minimum number of observations that would be required to obtain different Kappa values via the commonly used R package “irr” (Gamer et al., 2019). To avoid the main limitation of the clipping method (it can bias the second rater into assuming the presence of a codable behaviour; see above), we introduced into the data set some instances where no behaviour occurred, as a negative control to overcome this bias (‘dummy videos’). We used the “N2.cohen.kappa” function of the irr package where we specified the lower (unacceptable) Kappa value ( $\kappa_0 = .6$ ; Cohen, 1988) and the expected (see next paragraph for explanation of how this expectation is formed) value ( $\kappa_1 > .60$ , Cohen, 1988). We also specified the proportional anticipated prevalence of the target behaviours occurring, e.g., how likely is it that the behaviours of interest will occur. Users of the BRAVO workflow should therefore calculate or estimate the expected proportions of each individual behaviour in their dataset beforehand. The “N2.cohen.kappa function” limits the number of categories to 10, meaning that if there are more behavioural types, agreement between raters should be calculated as a binary (i.e., both raters code the same behaviour [1] or not [0]). This provides a more conservative (larger) and therefore more reliable estimate of agreement (McHugh, 2012); see S2 for illustration of this with our data. For readers who do not wish to use R in their analyses, sample size calculators, providing the same results, are available via GitHub under CC licence 4.0 (Arifin, 2021a, 2021b).

If the behavioural categories were independent, it would be possible to gain cross-category reliability. However, this is not the case with our simulated data because the behaviours are nested within each other, meaning that videos from each behaviour need to be included in the subset analysed by the second rater (Table 3). This forces the researcher’s hand, as it requires more clips to be sampled to achieve the same (arbitrary) necessary power level (.8). For the abovementioned example (41 categories) and to obtain a Kappa value of .8 (as advised by Cohen, 1988) the sample size of the subset analysed by the second rater should be  $N = 105$ . This value is achieved by the function a 50/50 likelihood that each rater would identify a behaviour (chance, given the dummy variables) an *expected*  $\kappa$  (here expected  $\kappa = .78$ ; this should be based on prior experience with the rater or values from relevant literature; the former in this case: the expected Kappa in this case was an average of DN’s published Kappa values (Needle et al., 2017, 2020).

**Table 4**

*Required Sample Sizes to Achieve a Power Level of .8*

Variable	Number of categories	Required N
Behaviour	2 <sup>†</sup>	105
Sub-Category	9	74
Category	3	85
Handedness	3	83
Sociality	3	83
Time-start	Continuous	65
Duration	Continuous	65

<sup>†</sup> Fine-grained behavioural categories were collapsed into binary agree/disagree owing to category constraints in the irr package.

*Note.* To achieve these values, expected agreement ( $\kappa_1 = .78$ ) and threshold agreement ( $\kappa_0 = .6$ ) (see main text) remained constant and probabilities were extracted from the simulated dataset from S2 (see supplementary script for details). All categorical variables had a .5 probability of no behaviour added and other relative probabilities reduced by 50% to account for the ‘dummy variables’, described above.

When calculating the required number of clips necessary for IRR assessment, one should conduct this kind of power analysis for all of the variables to be analysed. This includes Kappa calculations for categorical variables (Cohen, 1968) and Intraclass Correlation Coefficient for continuous variables (ICC; Bartko, 1966). The largest number of videos should then be used to ensure the validity of the calculations across variables. In our simulated example (Table 3), 105 out of 700 clips including both positive and dummy videos should be coded by the second rater ( $N = 105$  in the example from Table 3). Although we advise against it, if dummy clips are not used (e.g., the ‘clipping method’ described in the introduction), the 0.5 probability should not be added in the sample size calculations and the 50% adjustment should not be applied, which would require that the entire required  $N$  is filled with “positive” clips, i.e., behaviours from the ethogram.

### **Content Selection of Sample Provided to Second Rater**

As mentioned above, the clipping method has the disadvantage that the first rater could (intentionally or not) cherry-pick clear cut examples of the target behaviours and omit those that even the main rater is not certain of (as a part of the “raters degree of freedom” problem; see above). This limitation can be addressed in two ways:

- a) Select specifically those questionable examples (for the most conservative IRR)
- b) Randomly select behaviours

In the interest of avoiding yet another potential bias, we advise the latter method and outline it below. However, for reference, ‘questionable examples’ are those

where the first rater was unsure of the classification. However, such uncertainty to some extent implies a sub-standard ethogram and further operationalisation might be required.

In order to select the subsample that is to be provided to the second rater, one should assign unique ID numbers to each behavioural observation. This can be done by simply attaching ascending numbers to each row of the datasheet (see S1; though note that a set of random, fully arbitrary numbers is even preferable). Next, to select the behaviours for the sample clips containing behaviours of interest, the number of clips required (i.e. the largest necessary value from power calculations above) should be used to generate, without replacement, a series of randomly selected IDs. These can then be used to subset the clips from the full videos, using their timestamps. Once the “positive” clips are generated, the same number of negative controls (dummy clips) need to be generated, e.g., containing resting behaviours or behaviours *not* in the ethogram and thus outside the scope of the study.

To select the dummy clips, the main rater can simply look at their coding sheet and select clips from the times *not* covered by behaviours of interest (or, ideally, randomise all these non-occurrence after giving them their own IDs). When ‘clipping’ these dummy clips, the researcher should ensure that they are roughly centred on the mean duration of the ‘positive’ clips that will be given to the second rater (see S1 for an example of how to do this). A record of which clips are ‘dummy’ clips should then be entered into the first rater’s reliability subset (the data set of events matching in length and IDs the coding sheet provided to the second rater to perform their behavioural classification). Entries for dummy and positive clips/events should be randomly ordered in the coding sheet provided to the second rater and labelled exclusively by their generated ID to avoid providing any content information (see examples in S2). The second rater should also be provided in addition to the coding sheet and clips with a copy of the ethogram. We advocate to communicate (in writing, on top of the ethogram) to the second rater that not all clips may contain codable behaviours. However, it is *imperative* that the rater is *not* told any information about the proportion of dummy clips to positive clips. This is to reduce respondent bias in favour of one or the other response (no response vs. behaviour code).

The second rater should not be given any other instruction outside of the information contained on the document containing the ethogram (if further instruction is needed, this would indicate a problem with the ethogram), beyond communicating that not all clips may contain a codable behaviour. This is imperative in ensuring that the second rater remains naïve to the aims of the study and in ensuring that process of IRR assessment does not bias its results.

In our simulations, the second-rater’s behavioural classification was generated based on the first rater’s data set by introducing random errors into the coding while maintaining an arbitrary 65% accuracy between the two raters. In the following

section we describe how these two data sets would be compared in order to derive IRR measures.

### IRR Measures

This section outlines the approaches used to get an IRR score for both categorical and continuous variables, to give a clearer workflow, the worked example is referenced throughout.

#### IRR for Categorical Variables

When assessing IRR of categorical variables in a two-rater system (as we used here), Cohen's Kappa should be used (McHugh, 2012). If the levels of a categorical variable are ordered, it is possible to weight responses, so disagreements are 'punished' depending on the degree. However, in most cases, the levels of categorical variables are not ordered and therefore 'standard' unweighted Kappa values should be applied. Depending on the software used, it might be required to transform the data into numeric variables or to reset value labels. For instance, if one were to send the datasheet to a second rater via CSV for use in Microsoft Excel then the Kappa values are to be calculated in SPSS the string variables (i.e., behaviour names) will not be appropriate. If using the R script which accompanies this article, there are integrated steps to preserve the formatting of the factors, thus allowing behavioural codes to be read into the main dataset (providing that they are identical, which is the advantage of the data entry form outlined in the script). Regardless of how this is achieved, the coding of the first and second raters should be compared (using the random behavioural ID as a reference point) in order to determine the degree of agreement between the raters.

In our simulations, the results of a Cohen's Kappa analysis revealed a moderate level (McHugh, 2012) of agreement between raters for all categorical variables (see Table 4). These values are above the acceptable level of agreement ( $\kappa > .60$ ; Cohen, 1968) and therefore can be considered 'reliable'. Note however that the level of agreement was predetermined in our simulations and that these results are simply to demonstrate correct reporting and calculation of the Kappa statistic.

**Table 5**

*Kappa Values for Each of the Categorical Variables in the Simulated Data Set*

Category	Kappa Value ( $\kappa$ )	Z	p-value
Sub-category	.68	9.71	< .001
Behaviour	.66	14.8	< .001
Handedness	.66	17.9	< .001
Sociality	.65	9.37	< .001
Sociality	.68	9.78	< .001

*Note.* All values are calculated using unweighted calculations, two simulated raters and 106 observations.

In cases where there is a fundamental disagreement between raters and IRR assessments do not reach acceptable levels of agreement; clear and relevant changes must be made to the behavioural ethogram. Researchers are required to update the ethogram if it is suspected to have been biased during its construction. The ethogram must also be revised if new behaviours have been observed during video coding, which might lead to discussions between the first and second raters – remembering that (due to the nature of this conversation) the second rater may no longer be considered naïve to the study, rendering them unsuitable as second raters. However, first and second raters should never “talk through” the discrepant cases, come to a *post-hoc* agreement between raters, and then run the changed dataset for a “new IRR”, as this artificially increases reliability at the expense of validity (and certainly such a practice should not go unreported).

First and second raters *could* discuss the study if their aim is to revise the ethogram for *future* use. In this case, one may seek to first understand the major points of disagreement through a confusion matrix and then ‘talk through’ the unclear cases. It is possible to generate a ‘confusion matrix’ (Table 6) using the ‘caret’ package in R (Kuhn, 2021), to determine the areas of the ethogram which are/are not at ‘fault’ through lower/higher balanced accuracy scores. The ‘confusionMatrix’ command also provides percentage agreement and Kappa statistics, but the output is much more in depth than that of kappa2 from ‘irr’. An example of this is performed in the R script (S2) and the confusion matrix can be found in Table 6. Once the ethogram is modified, the main rater needs to code all videos again, and the second (new) rater needs to be given a new set of selected videos to code.

Note that without making changes to the ethogram and recoding the videos, it would not be appropriate to blame the second rater, resample videos and ‘try again’ for reliability with the same or a different second rater, as this could be considered as ‘fishing’ (a questionable coder practice), a practice not consistent with the BRAVO workflow.

### **IRR for Continuous Variables**

As we alluded to in the introduction, it can be difficult to attain good reliability for continuous variables, such as start times, owing to the granularity of the measure, i.e., if one measures to the second, how specific does one need to be to obtain good reliability? To account for this limitation, we will now outline two approaches and demonstrate the differences in reliability scores obtained following both. The first approach assumes that the timestamp of the behaviour coded by both raters needs to be identical for the scores to be considered reliable – we will refer to this as the ‘*fine-grained*’ approach. This does not require any data manipulation, and the values should be simply compared using a one-way, single-unit, consistency ICC test. However, to apply ‘fine-grained’ coding one does need to specify the level of



**Table 6**  
*Confusion Matrix of Behavioural Sub-Categories*

		Main Rater								
		Play (.78)	Display (.83)	Groom (.90)	Courtship (.76)	Fruit- Hammer Foraging (.79)	Stick- Based Foraging (.87)	Anvil Smash (.90)	Rolling Pin Techniques (.82)	Dummy (.84)
S e c o n d R a t e r	Play	<b>4</b>	0	0	0	0	0	0	0	2
	Display	0	<b>6</b>	0	0	0	0	0	0	1
	Groom	0	2	<b>5</b>	0	1	0	0	0	0
	Courtship	1	0	0	<b>5</b>	0	0	0	0	3
	Fruit- Hammer Foraging	0	1	0	0	<b>3</b>	0	0	0	1
	Stick- Based Foraging	0	0	0	1	1	<b>7</b>	0	0	1
	Anvil Smash	0	0	0	0	0	0	<b>4</b>	1	0
	Rolling Pin Techniques	1	0	0	0	0	0	0	<b>2</b>	1
	Dummy	1	0	1	3	0	2	1	0	<b>44</b>

*Note.* Agreements are in bold and balanced accuracy of each sub-category is displayed in parentheses

precision with which times are coded (i.e., minutes, seconds, milliseconds, etc.) which can introduce subjectivity. The second approach (in our example) allows a five-second grace period around the timestamp for two ratings to be considered in agreement. Of course, this figure of five seconds is arbitrary and can be changed. The figure chosen should be based on the length of the target behaviours and detailed explicitly in the reports. We will call this the “*coarse-grained approach*”. It requires the data to be transformed into categorical agreements through a logical argument (see R script [S2] for example and execution; essentially grouping the agreements into agree, disagree-too early or disagree-too late). To do this transformation, it is necessary to create a measure of absolute difference between the times observed by two raters. A logic statement can then be used to determine if the second rater agreed with the main rater or not (see example in S2). As Kappa calculations require two sets of ratings to be compared, a second dataset needs to be generated that demonstrates agreement for those behaviours within the five second (or other arbitrary) ‘grace’ period (see S2 [start time example]).

For our simulated behaviours we conducted inter-rater reliability assessments for the continuous variables of ‘start\_time’ and ‘end\_time’. It was not necessary to compute inter-rater reliability scores for duration as it is a direct calculation from these variables. The results of an ICC test, where the values are correlated with one another and a ratio of variance of interest/total variance is computed (Liljequist et al., 2019), showed moderate inter-rater reliability (Koo & Li, 2016) in both continuous measures (ICCstart\_time = 0.68,  $F(105,106) = 5.28$ ,  $p < .001$ ; ICCend\_time = 0.68,

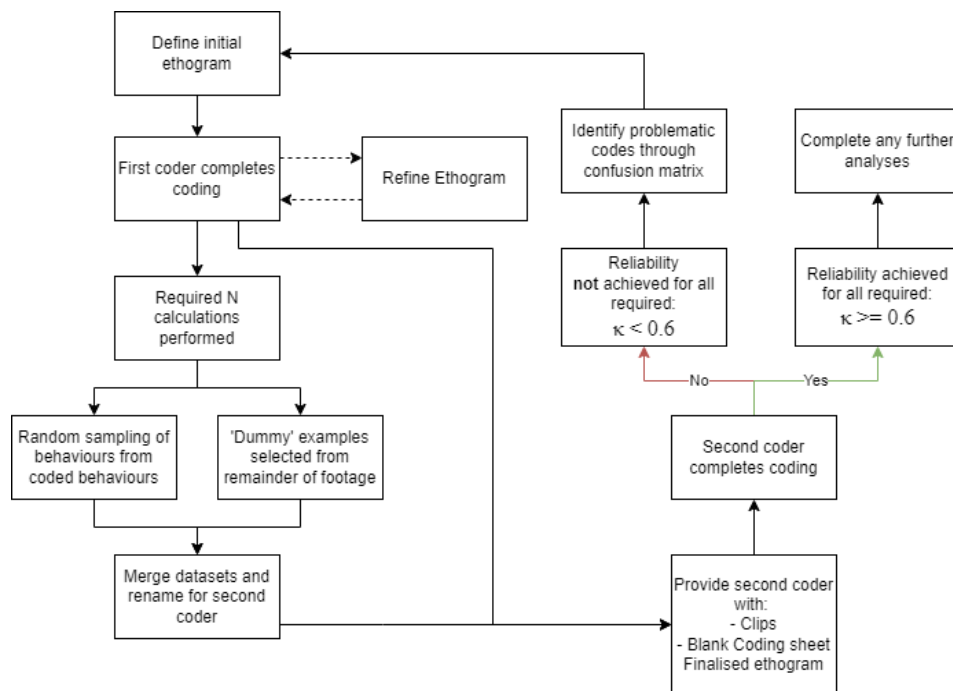
$F(105,106) = 5.28, p < .001$ ; note, as these are simulated by randomly sampling from the main simulated data set with replacement, the values are the same and are both included here for demonstration purposes).

Although it would be *possible*, it would be technically incorrect to calculate Kappa statistics to assess IRR of categorical variables as statistical software may treat the data as ordinal and this would lead to an artificially suppressed IRR value. ICC allows for some variance in continuous data in a way that Kappa cannot. As mentioned above, in cases where the behavioural duration is important, but the precision of the timings is less critical, it is possible to use a ‘coarse-grained’ approach where a margin for error is built into the IRR analysis to ensure that raters are not too ‘harshly punished’ for minor deviations which can be attributed to reaction time error or even software differences.

In the case of the example here, the coarse-grained method actually resulted in poorer IRR estimates ( $\kappa = .38, Z = 5.19, p < .001$ ) than the fine-grained approach. Yet, this might not always be the case. The utility of each method depends on the research question of the specific study. In this case, as the start, end and duration data were collected, it is possible to assume that they were pivotal and therefore the former fine-grained method (and ICC values) should be used.

**Figure 1**

*Flowchart Overview of the BRAVO Workflow*



**Table 7**

*Example Checklist for When Using the BRAVO Workflow*

Step	Description	Check
Define Initial Ethogram	All behaviours of interest to be included in the study are listed and described (use relevant literature where possible). Optional: Ideally example pictograms (or similar) are given.	<input type="checkbox"/>
	<b>Define structure of behaviour:</b>	
	Define beginning and end of behaviour.	<input type="checkbox"/>
	<b>Define use-case specific scenarios (list as required):</b>	
	What happens if a second behaviour happens simultaneously?	<input type="checkbox"/>
	What happens if an individual changes hands?	<input type="checkbox"/>
	What happens if an individual takes a break and continues doing the same action?	<input type="checkbox"/>
	When does a new event of the same action start?	<input type="checkbox"/>
	Is there a set limit to the length of a behaviour?	<input type="checkbox"/>
	Are behaviours to be grouped into 'bouts'? Should repetitive actions be grouped together? <i>E.g., should multiple strikes of a stone hammer on a nut be considered as one behaviour or multiple?</i>	<input type="checkbox"/>
Create Coding Sheet	<b>Coding sheet should contain:</b>	
	Every element of the behaviour to be coded in a separate column.	<input type="checkbox"/>
	One row per behaviour.	<input type="checkbox"/>
	Time stamps should be included in marking the start and end of each behaviour.	<input type="checkbox"/>
First Rater Completes Coding	The name of the first (main) rater should be included.	<input type="checkbox"/>
	Use ethogram to code all video footage for the behaviours listed in the ethogram.	<input type="checkbox"/>
Refine Ethogram (optional)	Behaviours added to the ethogram that are relevant to the aims of the study but are not present in the initial iteration. Should be added and defined when the behaviour is first observed.	<input type="checkbox"/>
	If required, go back and recode all clips coded already to include the new behaviour.	<input type="checkbox"/>
		<input type="checkbox"/>
Required <i>N</i> Calculations Performed	Use script in S2 to calculate how many clips are needed for the study.	<input type="checkbox"/>

Step	Description	Check
Random Sample Behaviours	Assign unique ID numbers to each behavioural observation.	<input type="checkbox"/>
	Generate the required number (from the previous step) of random clip IDs from the list of unique IDs.	<input type="checkbox"/>
	Subset the entire dataset for just the relevant clips.	<input type="checkbox"/>
Select 'Dummy' Examples (where applicable)	Use the completed code sheet to select and clip parts of videos not containing behaviours of interest.	<input type="checkbox"/>
	Ensure that 'dummy' clips are not a dissimilar in length to 'live' clips.	<input type="checkbox"/>
Merge and rename for second rater	Create a mapping table linking randomly generated IDs to the unique IDs assigned in the previous step.	<input type="checkbox"/>
	Make sure to record which randomly generated IDs relate to dummy clips.	<input type="checkbox"/>
Provision second rater	<b>Provide second rater with:</b>	<input type="checkbox"/>
	Blank coding sheet	<input type="checkbox"/>
	Clips of 'live' and 'dummy' behaviours Finalised ethogram	<input type="checkbox"/>
Perform reliability analysis	<b>Use an appropriate metric (see Konstantinidis et al., 2022 or Harvey, 2021):</b>	<input type="checkbox"/>
	Cohen's Kappa (2x raters, $N$ clips > 30, lowest count of instances of behaviours $\geq 5$ )	<input type="checkbox"/>
	Gwet's AC1 (2+ raters, $N > 30$ , lowest count of instances of behaviours $\leq 5$ )	<input type="checkbox"/>
	Fleiss' Kappa (3+ raters, $N$ clips > 30, lowest count of instances of behaviours $\geq 5$ )	<input type="checkbox"/>
	Determine acceptability of the result based on accepted thresholds from relevant literature.	<input type="checkbox"/>
Identify problematic codes and refine ethogram (optional)	Using the generated confusion matrix (see Table 1), identify 'problematic' codes (areas of substantial disagreement) and review the definition of behaviour.	<input type="checkbox"/>
	Consider providing more detailed descriptions and examples, if appropriate.	<input type="checkbox"/>
	May review the updated description with the previous second rater.	<input type="checkbox"/>
	If this step is reached (i.e., one round of IRR testing has 'failed'), then a second hypothesis naïve rater must be recruited.	<input type="checkbox"/>

*Note.* We are not claiming that this list is exhaustive. Boldface items are overarching steps containing nested sub steps. Italicised items are examples only and may require further population from a user.

## Conclusions and Discussion

Here, we present BRAVO (Balanced Reliability Assessment of Video Observations), a workflow for preparing and conducting interrater reliability assessments (IRR). This workflow is designed to limit currently wide-spread coder's degrees of freedom regarding the selection of video clips for IRR and to avoid unintentional questionable coder practices.

Our article is additionally intended to serve as a point of reference for the need to reconsider current practices, update reliability assessment methods generally and increase the transparency of IRR reporting. Reliability is not merely a box to tick. Instead, it is an important part of research in fields such as psychology, ethology and anthropology, and as such it requires to be done in ways that ensures the validity of behavioural classifications. As we have shown, current practices are seemingly maximised for convenience, at the expense of validity. We hope that our paper changes the perception of these practices.

Is a type of reliability assessment as we outline always required in behavioural sciences? Principally yes, though, as discussed in the introduction, there are cases where some variants are not possible e.g., field work in remote populations/areas or difficult conditions, might not permit video recording of data (though note that research that collects stable artefacts does require IRR; for the coding of these artefacts itself can be done repeatedly; e.g., applicable in primate archaeology and archaeology in general). As a case in point, some species of primates and birds spend the majority of time in the canopy of their rainforest habitat, it is difficult to get clear and/or sufficiently frequent video recordings in these cases and thus 'live' coding might need to suffice (also, such live videos can be very shaky and confusing). Yet, even here, two raters can be made present, to live code the same data (see work by Susan Perry and collaborators for the application of this procedure in field research; Perry, 1995; Perry et al., 2008). These cases still leave room for impromptu and difficult-to-report rounds of communication and thus convergence-by-communication, rather than independent reliability.

Essentially, our call here is for researchers to apply the same level of transparency that many now do with their inferential or exploratory analysis to the first step in many observational and behavioural studies. We suggest that by applying BRAVO (or the gold-standard of dual coding of all videos), it will be possible to improve the quality and consistency of the data collected and to further evolve the disciplines. This, along with other movements in open science (registered reports, preregistration, publishing of negative/null results, etc.) will aid us in distancing ourselves from the ongoing 'replicability crisis' (e.g., Open Science Collaboration, 2015; Wiggins & Christopherson, 2019) and preserve the validity of and public faith in our data and the conclusions drawn from these. In many (though not all) cases we can achieve validity in our conclusions only with reliable data.

Finally, we wish to once more make clear that we do not see BRAVO as the only possible improvement in reliability methods. Neither do we see BRAVO as an improvement across the board. We merely suggest that BRAVO is a stepping-stone along the way to improve reliability assessments. Many more improvements may be possible – and we are looking forward to seeing them in print and practice. For example, as one of the two reviewers of this article outlined, future approaches towards a best-possible practice could potentially include using naïve raters only, more than two raters, raters from different labs, more independent creations of ethograms, adaptive learning software guiding the coding and perhaps even artificial intelligence and machine learning (where these work reliably and validly). We agree, but we need to start somewhere. The BRAVO workflow is our contribution to this road and to this debate.

### References

- Acerbi, A., Snyder, W. D., & Tennie, C. (2022). The method of exclusion (still) cannot identify specific mechanisms of cultural inheritance. *Scientific Reports*, *12*(1), Article 21680. <https://doi.org/10.1038/s41598-022-25646-9>
- Allritz, M., McEwen, E. S., & Call, J. (2021). Chimpanzees (*Pan troglodytes*) show subtle signs of uncertainty when choices are more difficult. *Cognition*, *214*, Article 104766. <https://doi.org/10.1016/j.cognition.2021.104766>
- Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour*, *49*(3), 227–267. <https://www.jstor.org/stable/4533591>
- Arifin, W. N. (2021a). *Sample size calculator (web): ICC*. <https://wnarifin.github.io/ssc/ssicc.html>
- Arifin, W. N. (2021b). *Sample size calculator (web): Kappa*. <https://wnarifin.github.io/ssc/sskappa.html>
- Bakar, Y., Özdemir, Ö. C., Sevim, S., Duygu, E., Tuğral, A., & Sürmeli, M. (2017). Intra-observer and inter-observer reliability of leg circumference measurement among six observers: A single blinded randomized trial. *Journal of Medicine and Life*, *10*(3), 176–181. <https://pmc.ncbi.nlm.nih.gov/articles/PMC5652265/>
- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, *19*(1), 3–11. <https://doi.org/10.2466/pr0.1966.19.1.3>
- Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H. Jüni, P., & Cuijpers, P. (2016). Comparative efficacy of seven psychotherapeutic interventions for patients with depression: A network meta-analysis. *Focus*, *14*(2), 229–243. <https://doi.org/10.1176/appi.focus.140201>
- Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative Psychology*, *9*(1), 3–26. <https://doi.org/10.1037/qup0000196>

- Byrne, D. A. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, 56, 1391–1412. <https://doi.org/10.1007/s11135-021-01182-y>
- Cartmill, E., & Byrne, R. W. (2011). Addressing the problems of intentionality and granularity in non-human primate gesture. In G. Stam & M. Ishino (Eds.), *Integrating Gestures* (pp. 15–27). John Benjamins Publishing Company. <https://doi.org/10.1515/9783110668568>
- Charbonneau, M., & Bourrat, P. (2021). Fidelity and the grain problem in cultural evolution. *Synthese*, 199, 5815–5836. <https://doi.org/10.1007/s11229-021-03047-1>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220. <https://doi.org/10.1037/h0026256>
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Academic Press, Inc.
- Denis, C. M., Gelernter, J., Hart, A. B., & Kranzler, H. R. (2015). Inter-observer reliability of DSM-5 substance use disorders. *Drug and Alcohol Dependence*, 153, 229–235. <https://doi.org/10.1016/j.drugalcdep.2015.05.019>
- Di Eugenio, B. D., & Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics*, 30(1), 95–101. <https://doi.org/10.1162/089120104773633402>
- Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *Various coefficients of interrater reliability and agreement* (0.84.1). CRAN. <https://cran.r-project.org/web/packages/irr/index.html>
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34. <https://doi.org/10.20982/tqmp.08.1.p023>
- Harvey, N. D. (2021, December 4). *A simple guide to inter-rater, intra-rater and test-retest reliability for animal behaviour studies*. OSF Preprints. <https://doi.org/10.31219/osf.io/8stpy>
- Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., Ravaud, P., & Brorson, S. (2012). Observer bias in randomised clinical trials with binary outcomes: Systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ*, 344, Article e1119. <https://doi.org/10.1136/bmj.e1119>
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532. <https://doi.org/10.1177/0956797611430953>

- Konstantinidis, M., Le, L. W., & Gao, X. (2022). An empirical comparative assessment of inter-rater agreement of binary outcomes and multiple raters. *Symmetry*, *14*(2), Article 262. <https://doi.org/10.3390/sym14020262>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Koops, K., Furuichi, T., & Hashimoto, C. (2015). Chimpanzees and bonobos differ in intrinsic motivation for tool use. *Scientific Reports*, *5*, Article 11356. <https://doi.org/10.1038/srep11356>
- Kuhn, M. (2021). *caret: Classification and Regression Training* (6.0-88). <https://cran.r-project.org/web/packages/caret/caret.pdf>
- Liljequist, D., Elfving, B., & Roaldsen, K. S. (2019). Intraclass correlation – A discussion and demonstration of basic features. *PLoS ONE*, *14*(7), Article e0219854. <https://doi.org/10.1371/journal.pone.0219854>
- McHugh, M. L. (2012). Lessons in biostatistics interrater reliability: The kappa statistic. *Biochemica Medica*, *22*(3), 276–282. <https://pubmed.ncbi.nlm.nih.gov/articles/PMC3900052/>
- Mohan, V., Perri, M., Paungmali, A., Silitertpisan, P., Joseph, L. H., Jathin, R., Mustafa, M. B., & Nasir, S. H. B. M. (2017). Intra-rater and inter-rater reliability of total faulty breathing scale using visual observation and videogrammetry methods. *Journal of Bodywork and Movement Therapies*, *21*(3), 694–698. <https://doi.org/10.1016/j.jbmt.2016.10.007>
- Neadle, D., Allritz, M., & Tennie, C. (2017). Food cleaning in gorillas: Social learning is a possibility but not a necessity. *PLoS ONE*, *12*(12), Article e0188866. <https://doi.org/10.1371/journal.pone.0188866>
- Neadle, D., Bandini, E., & Tennie, C. (2020). Testing the individual and social learning abilities of task-naïve captive chimpanzees (*Pan troglodytes* sp.) in a nut-cracking task. *PeerJ*, *8*, Article e8734. <https://doi.org/10.7717/peerj.8734>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. <https://doi.org/10.1126/science.aac4716>
- Perry, S. (1995). *Social relationships in wild white-faced capuchin monkeys, Cebus capucinus* [Doctoral dissertation, The University of Michigan]. ProQuest Dissertations & Theses. <https://www.proquest.com/docview/304200702?sourcetype=Dissertations%20&%20Theses>
- Perry, S., Manson, J. H., Muniz, L., Gros-Louis, J., & Vigilant, L. (2008). Kin-biased social behaviour in wild adult female white-faced capuchins, *Cebus capucinus*. *Animal Behaviour*, *76*(1), 187–199. <https://doi.org/10.1016/j.anbehav.2008.01.020>



- Roberts, K., Dowell, A., & Nie, J. B. (2019). Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. *BMC Medical Research Methodology*, *19*(1), Article 66. <https://doi.org/10.1186/s12874-019-0707-y>
- Sainani, K. L. (2017). Reliability statistics. *PM&R*, *9*(6), 622–628. <https://doi.org/10.1016/j.pmrj.2017.05.001>
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Tecwyn, E. C., Mazumder, P., & Buchsbaum, D. (2023). One-and two-year-olds grasp that causes must precede their effects. *Developmental Psychology*, *59*(8), 1519–1531. <https://doi.org/10.1037/dev0001551>
- Van Leeuwen, E. J., Staes, N., Brooker, J. S., Kordon, S., Nolte, S., Clay, Z., Eens, M., & Stevens, J. M. (2023). Group-specific expressions of co-feeding tolerance in bonobos and chimpanzees preclude dichotomous species generalizations. *iScience*, *26*(12), Article 108528. <https://doi.org/10.1016/j.isci.2023.108528>
- Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202–217. <https://doi.org/10.1037/teo0000186>
- Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, *13*, Article 61. <https://doi.org/10.1186/1471-2288-13-61>

Received: August 12, 2024

