





# Inter-Rater Agreement, Data Reliability, and The Crisis of Confidence in Psychological Research

Cathryn M. Button<sup>a</sup>, Brent Snook<sup>a</sup>   & Malcolm J. Grant<sup>a</sup>

<sup>a</sup>Memorial University

**Abstract** ■ In response to the crisis of confidence in psychology, a plethora of solutions have been proposed to improve the way research is conducted (e.g., increasing statistical power, focusing on confidence intervals, enhancing the disclosure of methods). One area that has received little attention is the reliability of data. We note that while it is well understood that reliability of measures is essential to replicability, there is a failure to apply some measure of data reliability consistently, or to correct for chance when assessing agreement. We discuss the problem of relying on Percent Agreement between observers as a measure of reliability and describe a dilemma that researchers encounter when assessing contradictory indicators of reliability. We conclude with some pedagogical strategies that might make the need for reliability measures and chance correction more likely to be understood and implemented. By so doing, researchers can contribute to solving some aspects of the crisis of confidence in psychological research.

**Keywords** ■ reliability; inter-rater agreement; Kappa; Percent Agreement; research methods, confidence intervals.

**Acting Editor** ■ Denis Cousineau (Université d'Ottawa)

 [bsnook@mun.ca](mailto:bsnook@mun.ca)

 [10.20982/tqmp.16.5.p467](https://doi.org/10.20982/tqmp.16.5.p467)

## Introduction

Much has been written recently about the crisis of confidence in psychological research (Open Science Collaboration, 2015). Some scholars (e.g. Hawkins et al., 2018; Maxwell, Lau, & Howard, 2015) have argued that the lack of properly conducted replications is a major contributor to the crisis. Other scholars (e.g. John, Loewenstein, & Prelec, 2012; Schmidt & Oh, 2016) have argued that there are many replications, as evidenced by the numerous meta-analyses that have been published, and that the crisis of confidence in psychological research is due to either publication bias or questionable research practices. Further still, scholars (Feng, 2014; Lilienfeld, 2017) have argued that – regardless of whether it is publication bias or questionable research practices – the true cause of those issues pertains to either fraud, incompetence, or unconscious bias from the pressures associated with the aphorism “publish or perish”; all of which are alarming.

A plethora of solutions have been proposed to improve the way psychologists conduct their research. Increasing statistical power, focusing on confidence intervals, enhanc-

ing the disclosure of methods, and pre-registering predictions and methods have all been suggested (e.g. Asendorpf et al., 2013; Wilkinson & the Task Force on Statistical Inference, 1999). Although the need for greater attention to such research design and analysis issues has been raised, one area that has received little attention is the reliability of data. In the current paper, we focus on reliability as indicated by inter-rater agreement. Specifically, we discuss the failure of researchers to measure agreement and to do so appropriately, present a data reliability dilemma that researchers often encounter, and suggest some pedagogical strategies that might make researchers pay closer attention to data reliability and, thus contribute to resolving some aspects of the crisis of confidence in psychological research.

## (Dis)Agreement on Inter-Rater Agreement

A common problem in many research situations occurs when a dependent variable has aspects that are unavoidably subjective and thus potentially susceptible to a high degree of measurement error (i.e., unreliability; Cousineau, 2020). The measurement of subjective vari-



ables, such as those involving a judgment of some kind, is a widespread practice in psychology and other social sciences. For example, researchers may be interested in the occurrence of aggression exhibited by children in a school yard (Pepler & Craig, 1995) or whether persons engaged in group discussion show a particular leadership style (Larson, Foster-Fishman, & Franz, 1998). Typically, in this kind of situation researchers attempt to assess reliability by examining agreement between the judgments of two or more observers. A variety of measures of agreement between observers have been proposed and assessed (Feng, 2014; Hallgren, 2012; Zhao, Liu, & Deng, 2013). There has also been extensive discussion of the problems inherent in virtually every measure of inter-rater reliability and there appears to be little agreement on which measure a researcher should use when assessing data reliability.

A recent paper by Grant, Button, and Snook (2017) suggested one resolution of this debate. In a series of Monte Carlo simulations, Grant and his colleagues used a novel criterion,  $d$ -prime, to assess the performance of five reliability measure indices. The criterion  $d$ -prime is an unbiased indicator of raters' sensitivity to the true presence or absence of the characteristic being judged, and therefore is a good proxy for objective reality. The researchers found that Phi and Kappa coefficients performed best across variations in characteristic prevalence, and raters' expertise and bias. They also found that correlations with  $d$ -prime for Percent Agreement, Scott's Pi, and Gwet's AC1 were much lower. Grant and colleagues concluded that, in situations where two raters make a series of binary decisions, researchers should choose Phi or Kappa to assess inter-rater agreement because those indices were the least influenced by variations in the decision environment and characteristics of the decision makers. A further advantage of the Kappa statistic is that its standard error is known and thus confidence limits around any given value can be easily calculated (see McHugh, 2012). Adding confidence intervals around estimates of reliability provides an indication of the measure's precision; that is, how close the estimate of reliability is to the true reliability value (see Cumming & Finch, 2005).

Although methodologists disagree on the most appropriate measure, or in some cases, even on how to define reliability (Feng, 2015; Krippendorff, 2016; Zhao, Feng, Liu, & Deng, 2018), they do agree on the need to adjust for chance when computing inter-rater agreement. Over half a century ago, Cohen (1960) drew attention to this need, noting the problems inherent in simply using Percent Agreement as a reliability index. Specifically, Cohen noted "The most primitive approach has been to simply count up the proportion of cases in which the judges agreed... and let the issue rest there... It takes relatively little in the way of sophis-

tication to appreciate the inadequacy of this solution." (p. 38). Yet, in the ensuing years, Percent Agreement has been among the most popular measures in the published literature. For instance, Fallon (2017) assessed the reliability of data reported in six forensic psychology journals between 1974 and 2015. Of 291 studies that contained a subjective variable, almost half (47%) failed to report any measure of reliability, and of those that did, 25% were uncorrected for chance. Overall, it was estimated that 60% of studies in those forensic psychology journals that contained a subjective variable reported data of questionable reliability. Similarly, the results of an analysis of inter-coder reliability practices of all studies in two communication journals over thirty years found that Percent Agreement was the most frequently observed measure (23%) and was most consistently used over the period of observation (Feng, 2014). Feng concluded that the widespread reliance on Percent Agreement is troubling and described the problems with not correcting for chance agreement.

There is, then, a puzzling disconnect between advice from methods experts and actual research practice. But why? Several explanations are possible. One is that the advice of methods experts is confusing. It is not surprising to us that practicing researchers exhibit confusion over what reliability measures to use and how to interpret reliability indices because there appears to be disagreement among methodologists about what measure best adjusts for chance agreement. For instance, there are over 20 published indices of reliability, all of them based on different assumptions about the role played by chance (Feng, 2014; Krippendorff, 2004; Zhao et al., 2018). A second explanation for the popularity of Percent Agreement is that it is easy to calculate and widely understood. Most people have been calculating and interpreting percentages since middle-school. A third explanation is that Percent Agreement matches our intuitive understanding of what it means to agree on something, in a way that probabilistic thinking does not (Gigerenzer & Hoffrage, 1995; also see Tversky & Kahneman, 1974).

### An Inter-Rater Reliability Dilemma

How do we go about ensuring then that researchers adjust for chance agreement when checking data reliability? The first step is to ensure that researchers understand the importance of correcting for chance, and how to do it (not just clicking a button that mindlessly produces a value). In order to address this question, it may be helpful to consider some of the reasons why two observers might agree in their judgments. Consider two judges, Chris and Laura, who observe 100 children at play and code instances of cooperation. Each of them codes whether or not a child cooperates during a fixed time period. In that scenario, agree-



**Figure 1 ■** Matrices for Chris and Laura’s coding decisions regarding the presence and absence of cooperation by children in a playground. In Panel a, Chris and Laura say “yes” four times more often than “no”. In Panel b, Chris and Laura say “yes” and “no” equally often. In Panel c, both Chris and Laura say “yes” more often than “no” but differ in the ratio.

		Laura		
		Yes	No	Total
Chris	Yes	70	10	80
	No	10	10	20
Total		80	20	100

		Laura		
		Yes	No	Total
Chris	Yes	40	10	50
	No	10	40	50
Total		50	50	100

		Laura		
		Yes	No	Total
Chris	Yes	60	20	80
	No	0	20	20
Total		60	40	100

ment might occur because both Chris and Laura are actually able to code the true presence or absence of cooperation. Grant et al. (2017) referred to this kind of sensitivity as Observer Expertise, and it is this factor that is central to reliability. Second, agreement might occur because both Chris and Laura come to the task with similar assumptions about the prevalence of cooperation. Grant et al. (2017) called this factor Observer Bias. Finally, agreement might occur purely by chance; that is, Chris and Laura could have been blindfolded and agreement would still be high. It should be clear that bias and chance, although unrelated to true reliability, might contribute to an inflated Percent Agreement in many circumstances.

With those concepts in mind, let us work through a scenario faced by Chris and Laura. After making their observations independently, they transfer each of their 100 binary decisions side by side onto a spreadsheet. They then compute how often they agree on their observations with Percent Agreement and Kappa. They discover that Percent Agreement is 80% (proportion is .80) and the Kappa value is .38 [95% CI: 0.13, 0.62], see Figure 1, Panel a. As researchers, they are confronted with seemingly conflicting evidence as to the reliability of their coding. On the one hand, 80% agreement suggests to them that agreement is high for this task. Moreover, such a positive result (i.e., they interpret their data as reliable) may facilitate the publication of any subsequent findings. On the other hand, a Kappa value of .38 (and might be as low as .13) is well below the generally accepted level of .70 (see Landis & Koch, 1977). Note that the upper limit of the 95% CI is still below the generally accepted level of .70. Such a negative result means their data are unreliable and they need to rethink their coding.

In such a situation, Chris and Laura have at least two options to choose from. First, they could simply focus on Percent Agreement and forge ahead, while rationalizing that Kappa is somehow inappropriate for their situation. Alternatively, they could re-evaluate their coding system, for example, by clarifying how they define “cooperation”, and repeating the entire process by getting two new

people to complete the coding task, and computing Kappa again (while crossing their fingers that an acceptable value emerges). As one can envision, the latter option is obviously more effortful than the first (i.e., repeating the coding process, perhaps more than once) and the outcome more uncertain (it is not guaranteed that a higher Kappa can be achieved). The decision made here by the researcher is consequential as it ultimately determines whether the data are treated as spurious and the results worthless, or the data are reliable and worth sharing publicly.

### Stop Rolling the Dice: We Need to Correct for Chance

Although the decision to focus on Percent Agreement may be due to publication pressures, we also think that part of the decision to rely on Percent Agreement is likely due to a failure to consider the important role of chance agreement. We suspect that working through the process of calculating chance agreement would be instructive. Consider the data shown in Figure 1, Panel a. Both Chris and Laura think cooperation occurred for 80% of the children. The chance that they will both say “yes” that cooperation occurred is:  $80 \times 80 \div 100$ , or 64%. Similarly, the chance that they will both say “no” that cooperation did not occur is:  $20 \times 20 \div 100$ , or 4%. The chance that Chris and Laura will agree on either Yes or No is 68% (64% + 4%). Put simply, we would expect Chris and Laura to agree 68% of the time if they were blindfolded during the task (i.e., the level of agreement expected by chance). If expected agreement by chance (68%) is subtracted from actual agreement (80%), the true level of agreement is a mere 12%. After computing the adjusted index of Percent Agreement (12%) and Kappa (.38), Chris and Laura should no longer be confronted with conflicting evidence as to the reliability of their coding. Both values indicate that their observations were unreliable. By taking the role of expected agreement into account when using Percent Agreement, researchers will have a more accurate measure of reliability.

The aforementioned example assumes that the two coders agree about the base-rates regarding the frequency of cooperation by children and how far their base rates de-



part from 50-50. What happens when the base-rates and/or the ratio of yes/no decisions are different for each coder? That is, there are instances where the two observers come with different expectations about the prevalence of cooperation (i.e., their biases differ). Consider Panels a and b in Figure 1 showing the agreement and disagreement frequencies of Chris and Laura judging 100 cases when they either disagree about the base-rates, the ratio of yes/no decisions, or both. Note that, in both cases, the two observers still agree 80% of the time.

As can be seen in Figure 1, Panel b, both Chris and Laura think cooperation occurred for 50% of the children. The chance that they will both say “yes” that cooperation occurred is:  $50 \times 50 \div 100$ , or 25%. Similarly, the chance that they will both say “no” that cooperation did not occur is:  $50 \times 50 \div 100$ , or 25%. When the chance that Chris and Laura will agree on either Yes or No is added (25% + 25%), the final level of chance agreement is 50%. If chance agreement (50%) is subtracted from actual agreement (80%), the adjusted level of Percent Agreement is 30%; the Kappa value is .60 [95% CI: 0.44, 0.76]. As above, the discrepancy between the Kappa value and Percent Agreement is greatly reduced and the dilemma for the researchers is resolved.

In Figure 1, Panel c, both Chris and Laura say “Yes” more often than “No” but Chris is more extreme in the ratio of yes to no decisions (i.e., 4 to 1) than Laura (i.e., 3 to 2). Using the formulas above, the Percent Agreement by chance would be 56%. Thus, the adjusted Percent Agreement would be 80% minus 56%, or 24%. The value for Kappa in this case is .55 [95% CI: 0.36, 0.72], reflecting a slightly more severe but still modest correction for chance. Note that in all three examples, the confidence intervals exclude zero, which is suggestive of above-chance agreement.

### Final Thoughts

How can one replicate results if data are unreliable? You cannot. We suspect that if researchers categorized studies by the type of reliability measure reported (e.g., none, uncorrected for chance, or corrected for chance), the success of replication would be greatest for those that reported measures that corrected for chance. It is axiomatic that researchers should report some index of observer agreement when data involve judgments of any kind. Further, the measure of agreement needs to be corrected for chance (e.g., correlation, Kappa) and be reported with associated confidence intervals. Percent Agreement’s intuitive appeal does not outweigh the need for chance correction or for the need for confidence intervals.

A sea change is needed with regard to the attention researchers pay to the reliability of their data. Journal editors must call upon researchers to include a chance-

corrected measure of reliability and associated confidence intervals whenever research variables involve a judgment. Graduate student teachers and mentors must demonstrate the importance of such a practice and teach the mechanics of doing so. We recall not so long ago that reporting confidence intervals and effect sizes was the exception rather than the rule. Currently, their reporting has become standard in the research literature. We believe the same must happen to data reliability measurement if researchers are to produce trustworthy research.

### References

- Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108–119. doi:10.1002/per.1919
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46. doi:10.1177/001316446002000104
- Cousineau, D. (2020). How many decimals? Rounding descriptive and inferential statistics based on measurement precision. *Journal of Mathematical Psychology, 97*, 1–43. doi:10.1016/j.jmp.2020.102362
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American psychologist, 60*, 170–180. doi:10.1037/0003-066X.60.2.170
- Fallon, L. (2017). *An examination of the use of inter-rater reliability in forensic psychology journals (unpublished master’s thesis)*. NL: Memorial University. St. John’s.
- Feng, G. C. (2014). Intercoder reliability indices: Disuse, misuse, and abuse. *Quality & Quantity: International Journal of Methodology, 48*, 1803–1815. doi:10.1007/s11135-013-9956-8
- Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 11*, 13–22. doi:10.1027/1614-2241/a000086
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704. doi:10.1037/0033-295X.102.4.684
- Grant, M. J., Button, C. M., & Snook, B. (2017). An evaluation of interrater reliability measures on binary tasks using d-prime. *Applied Psychological Measurement, 41*, 264–276. doi:10.1177/0146621616684584
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in quantitative methods for psychology, 8*, 23–34. doi:10.20982/tqmp.08.1.p023



- Hawkins, R. X. D., Smith, E. N., Au, C., Arias, J. M., Catapano, R., Hermann, E., ... Frank, M. (2018). Improving the replicability of psychological science through pedagogy. *Psychological Science*, *1*, 7–18. doi:[10.1177/2515245917740427](https://doi.org/10.1177/2515245917740427)
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:[10.1177/0956797611430953](https://doi.org/10.1177/0956797611430953)
- Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*, 411–433. doi:[10.1111/j.1468-2958.3042004.tb00738.x](https://doi.org/10.1111/j.1468-2958.3042004.tb00738.x)
- Krippendorff, K. (2016). Misunderstanding reliability. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *12*, 139–144. doi:[10.1027/1614-2241/a000119](https://doi.org/10.1027/1614-2241/a000119)
- Landis, R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. doi:[10.2307/2529310](https://doi.org/10.2307/2529310)
- Larson, J. R., Jr., Foster-Fishman, P. G., & Franz, T. M. (1998). Leadership style and the discussion of shared and unshared information in decision-making groups. *Personality and Social Psychology Bulletin*, *24*, 482–495. doi:[10.1177/0146167298245004](https://doi.org/10.1177/0146167298245004)
- Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science*, *12*, 660–664. doi:[10.1177/1745691616687745](https://doi.org/10.1177/1745691616687745)
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? *What does "failure to replicate" really mean?* *American Psychologist*, *70*, 487–498. doi:[10.1037/a0039400](https://doi.org/10.1037/a0039400)
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*, 276–282. doi:[10.11613/BM.2012.031](https://doi.org/10.11613/BM.2012.031)
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943–950. doi:[10.1126/science.aac4716](https://doi.org/10.1126/science.aac4716)
- Pepler, D. J., & Craig, W. M. (1995). A peek behind the fence: Naturalistic observations of aggressive children with remote audiovisual recording. *Developmental Psychology*, *31*, 548–553. doi:[10.1037/0012-1649.31.4.548](https://doi.org/10.1037/0012-1649.31.4.548)
- Schmidt, F. L., & Oh, I. S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology*, *4*, 32–37. doi:[10.1037/arc0000029](https://doi.org/10.1037/arc0000029)
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131. doi:[10.1126/science.185.4157.1124](https://doi.org/10.1126/science.185.4157.1124)
- Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. doi:[10.1037/0003-066X.54.8.594](https://doi.org/10.1037/0003-066X.54.8.594)
- Zhao, X., Feng, G. C., Liu, J. S., & Deng, K. (2018). We agreed to measure agreement – redefining reliability de-justifies krippendorff's alpha. *China Media Research*, *14*, 1–15.
- Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind inter-coder reliability indices. In C. T. Salmon (Ed.), *Communication yearbook* (pp. 419–480). New York, NY: Taylor and Francis.

## Citation

Button, C. M., Snook, B., & Grant, M. J. (2020). Inter-rater agreement, data reliability, and the crisis of confidence in psychological research. *The Quantitative Methods for Psychology*, *16*(5), 467–471. doi:[10.20982/tqmp.16.5.p467](https://doi.org/10.20982/tqmp.16.5.p467)

Copyright © 2020, Button, Snook, and Grant. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 22/06/2020 ~ Accepted: 01/10/2020