General Article

Advances in Methods and Practices in Psychological Science
October-December 2024, Vol. 7, No. 4, pp. 1–20

DOI: 10.1177/25152459241275217
www.psychologicalscience.org/AMPPS

Validity and Transparency in Quantifying Open-Ended Data

Clare Conry-Murray[1], Tal Waltzer[2], Fiona C. DeBernardi[3], Jessica L. Fossum[4], Simona Haasova[5], Michael S. Matthews[6], Aoife O'Mahony[7], David Moreau[8], Myriam A. Baum[9], Veli-Matti Karhulahti[10], Randy J. McCarthy[11], Helena M. Paterson[12], Kara McSweeney[13], and Mahmoud M. Elsherif[14]

[1]Department of Health, Behavior, and Society, Johns Hopkins University, Baltimore, Maryland; [2]Psychology Department, University of California, San Diego, San Diego, California; [3]Psychology Department, University of Oregon, Eugene, Oregon; [4]School of Psychology, Family, and Community, Seattle Pacific University, Seattle, Washington; [5]Department of Marketing, University of Lausanne, Lausanne, Switzerland; [6]Special Education & Child Development, University of North Carolina at Charlotte, Charlotte, North Carolina; [7]School of Psychology, Cardiff University, Cardiff, Wales; [8]Centre for Brain Research and School of Psychology, University of Auckland, Auckland, New Zealand; [9]Psychology Department, Saarland University, Saarland, Germany; [10]Department of Music, Art and Culture Studies, University of Jyväskylä, Jyväskylä, Finland; [11]Department of Music, Art and Culture Studies, Northern Illinois University, DeKalb, Illinois; [12]Psychology Department, University of Glasgow, Glasgow, Scotland; [13]Psychology Department, Saint Joseph's University, Philadelphia, Pennsylvania; and [14]Department of Psychology, University of Birmingham, Leicester, England

Abstract

Quantitatively coding open-ended data (e.g., from videos, interviews) can be a rich source of information in psychological research, but reporting practices vary substantially. We provide strategies for improving validity and reliability of coding open-ended data and investigate questionable research practices in this area. First, we systematically examined articles in four top psychology journals ($N = 956$) and found that 21% included open-ended data coded by humans. However, only about one-third of those articles reported sufficient details to replicate or evaluate the validity of the coding process. Next, we propose multiphase guidelines for transparently reporting on the quantitative coding of open-ended data, informed by concerns with replicability, content validity, and statistical validity. The first phase involves research design, including selecting data and identifying units reliably. The second phase includes developing a coding manual and training coders. The final phase outlines how to establish reliability. As part of this phase, we used data simulations to examine a common statistic for testing reliability on open-ended data, Cohen's κ, and found that it can become inflated when researchers repeatedly test interrater reliability or manipulate categories, such as by including a missing-data category. Finally, to facilitate transparent and valid coding of open-ended data, we provide a preregistration template that reflects these guidelines. All of the guidelines and resources provided in this article can be adapted for different types of studies, depending on context.

Keywords

coding, data analysis, data coding, open science, psychology, social sciences

Received 11/17/23; Revision accepted 7/19/24

The open-science movement has raised critical questions about how to conduct research more transparently in the psychological sciences (Hales et al., 2019), with the ultimate goal of improving validity in measurement,

Corresponding Author:
Clare Conry-Murray, Department of Health, Behavior, and Society, Johns Hopkins University, Baltimore, Maryland
Email: clareconrymurray@gmail.com

statistics, and the conclusions that are drawn (Vazire et al., 2022). Most efforts to improve psychological research have focused on quantitative data, but recently, qualitative research has also received more attention (see e.g., Campbell et al., 2023; Steltenpohl et al., 2023). The focus of the current article is the quantitative coding of open-ended data, which includes elements of both quantitative and qualitative analysis. Open-ended data are data that are not constrained by predefined responses. Examples of open-ended data include text-based free-response survey formats, interviews, images, social media postings, video materials, and observations of live behavior. Open-ended data are quantified when humans assign codes to summarize the data for statistical analysis. As part of this process, transparent research practices can help researchers and readers evaluate whether the codes and the conclusions drawn from the quantitative analyses accurately reflect the original meaning of the data.

There are many advantages to using open-ended data. Open-ended data allow participants to behave or provide responses with fewer limitations than forced-choice formats that have predetermined options (Adler et al., 2017; Braun & Clarke, 2013; Howitt & Cramer, 2011; Reja et al., 2003). Text, images, and behaviors can all be coded. Open-ended text or interviews can allow participants to provide "meanings through and in their own words" (Ruona, 2005, p. 234), which can give insight into participants' diverse reasons or motivations for thinking and behavior. The different types of open-ended data can provide insight into cognition or behavior with fewer restrictions imposed by researchers. This means that open-ended data can be especially useful when research is conducted in new settings, when findings are complex or context-specific, or when potential future hypotheses need to be developed. Data from open-ended survey responses can also be used in conjunction with forced-choice methods, either to offer alternative explanations to what was found via closed-ended measures (Jackson & Trochim, 2002) or to establish the validity of closed-ended questions (Singer & Couper, 2017).

Quantitative coding differs from some qualitative methods, such as reflexive thematic analysis (e.g., Braun & Clarke, 2013) because it allows researchers to assess the frequency of derived themes. When codes are well defined and agreement is established, it may also be possible to replicate findings or to discover patterns that generalize to other contexts.

However, there are dangers to the process of quantifying open-ended data. Quantification replaces rich text or other data with categories that summarize their meaning from the researchers' perspectives. The effect is that quantification can result in a loss of the specific meaning participants attempted to convey (Braun & Clarke, 2013). At the same time, quantifying open-ended data and indicating that the coding is reliable through agreement metrics can give the illusion of objectivity (see e.g., Natow, 2022) when, in fact, quantification is still subject to bias. When coding is conducted without transparency about the process, this problem is magnified because a lack of transparency can hide bias and other problems that undermine validity (Linneberg & Korsgaard, 2019; Syed & Nelson, 2015).

Across the field of psychology, there has been an effort to increase transparency of methods, materials, and analyses to examine studies' validity and potentially to replicate findings (Nosek et al., 2022). Recent studies suggest that transparent research practices may help improve research quality. For example, preregistration can increase power and sample size (Tenney et al., 2021; van den Akker et al., 2024). Initial evidence also suggests that Registered Reports (i.e., in which the research design and hypotheses are evaluated before data collection) show signs of increased quality compared with similar articles from a more traditional publishing model (Soderberg et al., 2021). Transparent research practices have only recently started to be adopted, and transparency is still lacking across the field of psychology (see e.g., Aguinis & Solarino, 2019; López-Nicolás et al., 2022).

The quantitative coding of open-ended data relies on researchers to code with fidelity to participants' original meaning using logical inference and appropriate statistics, but as we show below, much of that process is done with very little transparency. To aid in the movement toward a more transparent and valid psychological science, we developed a methodological framework to guide researchers and editors who work with quantitative coding of open-ended data.

In this article, we first report on common practices in how researchers construct quantitative summaries from open-ended data. We then describe strategies for increasing transparency in reporting the process of coding open-ended data, which can make the process more reliable and valid. We next use data simulations to examine how a lack of transparency in reporting metrics of agreement can conceal what we call "kappa-hacking," or misleading reports of interrater reliability. Finally, we provide a preregistration template and guidelines for researchers, editors, and reviewers that can be adapted to the needs of different types of studies using open-ended data to make it easier to use transparent and valid practices. The information provided in this article offers a starting point for discussions about best practices in quantifying open-ended data but no rigid rules or requirements. These considerations will vary for different types of research and must be taken in context.

The Current State of Reporting on Coding of Open-Ended Data: A Systematic Analysis of Recently Published Psychology Articles

To assess current practices in reporting on the coding of open-ended data in peer-reviewed publications, we analyzed academic articles from four high-impact psychology journals across different areas: *Cognitive Psychology*, *Developmental Science*, *Journal of Personality and Social Psychology*, and *Psychological Science*. Our aim was to select journals that are representative of different subareas of psychology and psychology as a whole. Although we do not consider the impact factor to serve as an accurate metric of scientific quality or impact, we chose journals with high impact factors (3.00 and above at the time of analysis) to pursue a sample of articles that are likely to be cited and read by researchers. Because of time and resource constraints, we restricted our analysis to articles published by the journals from the years 2020 to 2021. We examined all articles in every volume and issue published in those years and assessed how the coding of open-ended data was reported ($N = 956$ articles).

Methods: coding the content of published articles

For every article, we first assessed whether it contained any open-ended data, and if so, we coded whether the researchers used any type of human coding and whether they provided details about their coding approaches (see Table 1). We focused on elements of the coding process that were relevant to validity, transparency, and reliability in quantitative analysis of open-ended data.

To examine open-ended data and how they were coded in these articles, two of the authors (T. Waltzer and F. C. DeBernardi) discussed the goals of our article, drafted the proposed coding scheme, and conferred with

Table 1. Coding Categories and Agreement Scores (κ)

Category	Applied to	Description	κ
Article has open-ended data	Full data set ($N = 956$)	We coded an article as having open-ended data if it had a free-response format and if the data were reported or analyzed in the article.	.86
Researchers used coding	Articles with open-ended data ($n = 299$)	We coded an article as using coding if it involved having humans classify broader data into narrower categories and reported the results.	.88
Coding manuals provided	Articles with coding ($n = 200$)	We coded an article as having a coding manual if there were designated materials listing categories with some explanation.	.95
Coding manuals were detailed	Articles with coding manuals ($n = 72$)	We coded a manual as having sufficient detail if it included descriptions for most categories and provided examples (unless categories were self-evident).	1
Reliability information	Articles with coding ($n = 200$)	We coded an article as having information about reliability if it indicated what choice was made for agreement (e.g., interrater reliability, consensus).	.94
Extra materials	Articles with coding ($n = 200$)	We coded whether an article had public, additional materials outside the manuscript to supplement the main manuscript (e.g., OSF, GitHub, SOMs).	1
Positionality statements	Articles with coding ($n = 200$)	We coded an article as having a positionality statement if the authors described their identities or background somewhere in the article, regardless of whether it was related to the open-ended data.	N/A[a]

Note: SOMs = supplementary online materials.
[a]Too few instances of this code to calculate interrater reliability, but coders agreed 100% of the time.

> Three trained research assistants conducted the coding. For evaluations, the Cohen's kappa was between .92 and .94. For justifications, the kappa was between .71 and .74.
>
> (Conry-Murray & Turiel, 2012, p. 150)

Fig. 1. Example report on interrater reliability in the coding of open-ended data.

all of the authors to finalize the list of coding categories, which is detailed in Table 1. To code each category, we searched through all available materials for a given article (e.g., main text, OSF, other online supplementary materials). Each category was coded dichotomously as "present" or "absent."

We assessed our agreement in coding each of the categories by having the two authors independently code a random subset of 20% of the articles. After achieving agreement using Cohen's κ in our first test phase, one of the authors coded the remaining data. Further details about this analysis are available in our Supplementary Online Materials (SOM) on OSF (https://osf.io/du6gy/).

Results: patterns of (non)transparency in recent publications

About one-third of all the articles (31%, n = 299) contained some type of open-ended data (e.g., interviews, open responses, gaze patterns).[1] Most of the 299 articles with open-ended data applied some type of human coding to the data (67% of articles with open-ended data or 21% of all journal articles coded, n = 200). Articles counted as using coding if humans but not automated or algorithmic processes classified broader data into narrower categories. The vast majority of these methods were quantitative, although a few used qualitative approaches (e.g., thematic analysis).[2]

We looked more closely at those 200 articles that used coding. Most articles with coding provided the statistics they used to calculate agreement (79%) and had some supplementary materials beyond the main text (83%; e.g., stimuli or survey instruments). However, only about one-third of the articles that used coding had some type of coding manual (36%, n = 72) in either the text or supplementary materials. Even among the articles with manuals, only about two-thirds of them had sufficient detail, which was just 23% (n = 46) of all articles with coding. In addition, only one article contained a positionality statement (< 1%).

Our analysis suggests that there is a lack of transparency in reporting coding processes, especially regarding providing details about the coding schemes, coder instructions, and the authors' positionalities. Instead, authors typically reported only interrater reliability metrics without further elaboration. A typical report lists coding categories only by their labels and then reports a measure of agreement, similar to the example in Figure 1. Our analysis suggests that typical reports of coding open-ended data in journals do not contain sufficient information to reproduce the step-by-step process of coding open-ended data or to adequately evaluate its validity.

Suggested Guidelines for Transparency at Each Stage of the Process of Coding Open-Ended Data

What information is most essential to be included in reports of open-ended data that are quantitatively coded? There are many different methods of coding available (see e.g., Hallgren, 2012; Wicherts et al., 2016; and our SOM for terms, their definitions, and further reading: https://osf.io/du6gy/), and many decisions affect the validity and reliability of the results, similar to the garden of forking paths described by Gelman and Loken (2014). Coding open-ended data is further complicated by the fact that the specific details of coding are often based on tacit knowledge (Collins, 2001; Kiger & Varpio, 2020; Wilson, 2009). This means that each lab, research group, or individual researcher has their own set of methods and preconceived notions about their process, and this tacit knowledge is often opaque to others outside the research team.

In this section, we examine different considerations for coding open-ended data at each of three phases of the coding process: (a) research design and planning, (b) coding-manual development and training, and (c) establishing and reporting agreement. In the flowchart in Figure 2, we provide an illustration of this step-by-step approach to coding open-ended data. For each of these phases, we focus on creating transparent, reliable, and valid research.

Our guidelines are designed to meet the goals of producing valid science that leads to meaningful findings. However, we recognize that in implementing these goals, each study has different challenges that may mean compromises must be made. The features of the study, such as the needs of participants and the resources available, can mean that it is necessary to focus attention on different practices. For example, when working with open-ended data, full transparency can put participants' privacy at risk in ways that can be dangerous (see e.g., Campbell et al., 2023; Pownall et al., 2023). In addition, making open-ended data available for inspection or reuse may require significant resources, such as the time and financial resources needed for documentation, anonymization, and translation (Elman & Kapiszewski, 2014; Karhulahti, 2022).

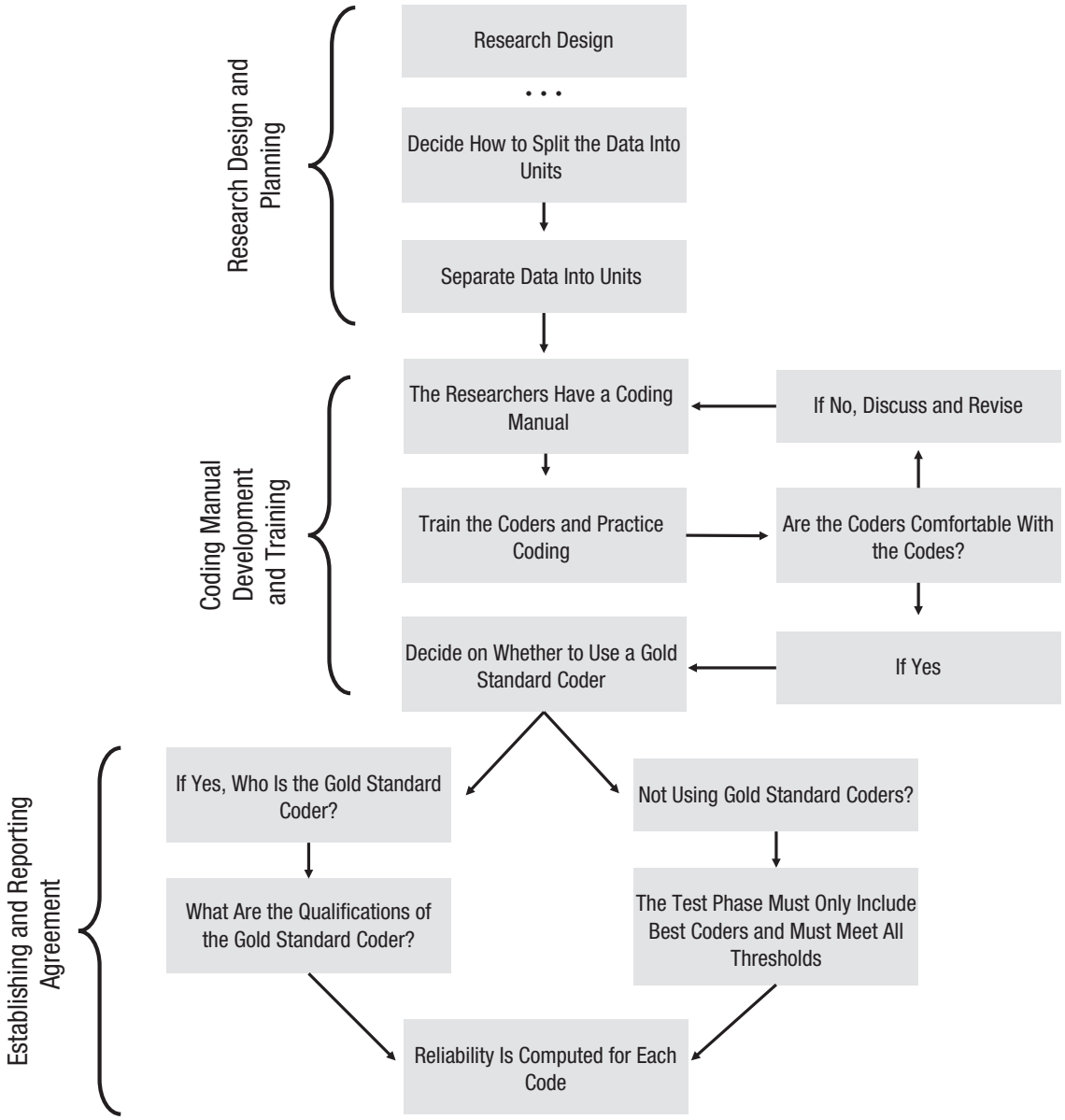

Fig. 2. A conceptual flowchart illustrating different steps in coding open-ended data.

We hope the guidelines that we propose in this article can be a start to deeper conversations around best practices in coding open-ended data and not the final word on any such conversations. Any suggestions in this article are not meant to be taken as absolute requirements without consideration of context. (And we note the risks of relying on default "checklists" for evaluating, say, journal-article submissions, which could bias publication rates against diverse forms of research that differ from the standard mold.) The following sections outline our proposed guidelines. Researchers can convey their efforts to make their work as reliable and transparent as possible by addressing the relevant areas described below, but when their research context is not suitable for full transparency, the researchers may choose to adjust their practices and explain their decisions (Adu, 2019; Peterson, 2019; Ryu, 2020).

Research design and planning

Goals and nature of the data: considerations in formulating a research design. Coding open-ended data is a resource-intensive process. Therefore, we recommend establishing and communicating a clear rationale for choosing quantitatively coded open-ended data over other methods of data collection (e.g., forced-choice questions) or analysis (e.g., producing rich narrative data for purely qualitative analysis). A clear rationale helps guide decisions

about whether and how to use quantitatively coded open-ended data.

The rationale for using open-ended data may differ depending on the type of data collected. If researchers plan to test hypotheses with open-ended data, they can specify those hypotheses ahead of time, for example, in a preregistration.

In their data collection, researchers can aid in assessments of validity by specifying the otherwise tacit details. For example, details about the context of interviews are important to document. Interviewers may have the opportunity for follow-up questions (Adler et al., 2017; Jackson & Trochim, 2002) if the context and study policies allow them. Interview responses can also differ depending on the level of privacy and confidentiality participants feel, including whether there are others who can hear their responses in the setting. When they are interviewed in settings with less privacy and researchers ask direct questions about sensitive issues, participants may be guarded in their responses (DeKeseredy et al., 2020; Grimm, 2010; Nederhof, 1985), or social-desirability effects may be increased (Fine & Elsbach, 2000). Written responses also have unique challenges that should be made explicit. The quality of independent written responses can be affected by details such as the size of the response box and the clarity of the instructions (Smyth, 2016). Open materials may convey some of these details (e.g., the size of the text box) but not others (e.g., who is in the lab when participants complete a survey). Researchers should be mindful of documenting the tacit factors in data collection that could affect the meaning of the data.

Selecting material to be coded. Another aspect of coding that sometimes remains tacit is the selection of materials to code. In some cases, all available material is coded, whereas in other cases, a subset of materials may be selected from a larger set of options. For example, for a study about content in children's television programs, it may not be possible to include all possible television programs. Therefore, choices must be made about which programs to include and how to sample them. Effort should be made to show that the material chosen is representative of the target population of materials. Alternatively, if researchers choose to focus on a nonrepresentative subset, they should provide a rationale for that choice and explain how their sample was chosen. Explanations should take into account a variety of factors, including research questions, the theory being tested, ethical considerations, and resource-related constraints.

Data preparation. Data preparation includes how and whether to transcribe audio files, de-identifying data, and other steps to prepare the data for coding. Details about how data are prepared for coding should be disclosed. For example, researchers may or may not choose to transcribe verbal data so that they can be coded more precisely. This decision and the quality of the transcription can have an impact on the meaning because a single missing word can transform the meaning of a phrase (Adler et al., 2017). De-identifying data must also be done with care because the context provided by identifying information may be important to the meaning of the data in some cases (e.g., Campbell et al., 2023). Because different research questions require different levels of transcription detail, researchers must assess what is optimal for their purposes; documentation that reflects both what was done and how these choices were made is helpful.

The unit of analysis. The selection of the unit of analysis is a part of data preparation that is sometimes overlooked. The unit of analysis (also referred to as a "segment"; Hruschka et al., 2004) is "the section of qualitative analysis that will receive a code" (Syed & Nelson, 2015, p. 377). Deciding what constitutes a unit for each study involves considering what unit is the most valid representation of the construct of interest and the practical implications of determining different units of analysis reliably. For example, a study of the Bechdel effect in television might have a sentence as a unit of analysis to see if women speak about topics unrelated to men in each sentence they speak, so each sentence is coded separately as a unit. Alternatively, the unit could be each turn talking, regardless of length of their speech, or the unit could be the time in minutes that each character speaks. In our analysis of coding in psychology journals from 2020 to 2021, the unit of analysis was a single published article.

In each of the cases above, it is relatively unambiguous to determine what constitutes a single unit. Perhaps because many units are unambiguous, many researchers do not report how the units were chosen. However, there are also many cases in which researchers make a subjective judgment about when a section of data constitutes a unit. For example, in participants' reasoning about their moral judgments, they may express multiple justifications (e.g., "That's not fair, and you could get in trouble for that!," in which the word "and" separates two different justifications, one related to fairness and one referring to punishment; for examples, see Waltzer et al., 2022, 2023). When units are subjective, establishing reliability for identifying the units separately from the reliability for coding the units is important. The process of identifying each unit of analysis can be strengthened if it is guided by clear criteria that are based on theory or clearly defined research questions, and documentation of this process can be included in the coding manual, described below.

Coding-manual development and training

Developing a coding manual. A coding manual is a document that details the definitions, rules, and procedures used to categorize units of open-ended data into codes (DeCuir-Gunby et al., 2011; Weston et al., 2001). A transparent and comprehensive coding manual is foundational for working with open-ended data because it defines how one reduces open-ended data into quantitative values (Adu, 2019; Weston et al., 2001; Wilson, 2009). Coding manuals are often too lengthy for a manuscript, but they can be shared in supplemental online materials. There are several considerations that should be addressed to make coding manuals more accessible and useful.

The specifics of choosing or developing a coding manual are subject to the researchers' goals, but validity should be prioritized as much as possible. When developing a new coding system adapted from existing work (e.g., Saldaña, 2016), it may be appropriate to use previous versions to show convergent, divergent, or incremental validity of the new coding system by comparing it with past systems (Adler et al., 2017). New coding systems can be developed as a bottom-up process, a top-down process, or a combination of both (Syed & Nelson, 2015), and these have different implications for validity. An example of a bottom-up process is when coders examine the responses and then propose a set of themes that reflects the data. In contrast, a top-down process starts with the themes relevant to the study's goals and examines whether they are actually present in the responses. When researchers create a new coding manual, they should consider what others may need to understand about the development of the coding categories in terms of how the category development may affect the meaning of the data.

As much as possible, valid coding should reflect the meaning intended by participants, ideally including the full spectrum of common responses in the data and in the population of interest (Syed & Nelson, 2015; Whittemore et al., 2001). For example, coding categories may be developed or examined from a random selection of 10% to 25% of the data, depending on frequency of the categories in the data (Adler et al., 2017), or categories may be derived from the entire data set, which helps to ensure that the codes created reflect the full range of responses (Syed & Nelson, 2015). One strategy to ensure validity, sometimes referred to as "member checking," involves consulting with participants for their feedback on the coding, interpretations, or conclusions of the research (Campbell et al., 2023; Lincoln & Guba, 1985). Documentation that details how data were used to develop the coding categories can help ensure that findings are interpreted correctly.

It is helpful to provide the set of rules defining whether a specific unit of analysis is or is not captured by a particular code, including both inclusion and exclusion criteria and examples (for more information on how to set these rules, see MacQueen et al., 1998). Documentation should also indicate whether codes are dichotomous, signifying absence or presence, or multivalue (e.g., low, medium, high; Hruschka et al., 2004) and how many codes can be used for each unit of analysis. Finally, any rules that transcend a single coding category, such as ranking different coding categories or hierarchical coding, should also be specified in documentation because some rules can affect the validity of the coding. For example, if coders are told that certain key phrases are indicative of a particular category or if they use rules such as "use x code as the default," they may assign items to a category even when the items do not conceptually fit. This can inflate the level of agreement at the expense of the validity of the findings. Therefore, this type of rule should also be documented.

Documentation should also explain when and where coding took place. Observational methods sometimes rely on coding that is done directly during data collection (on site) rather than afterward using recorded materials, such as video or audio (Sussman, 2016). Depending on the research, on-site coding may be appropriate, or it may lead to errors. When resources, such as time and money, do not allow for recording and transcription of data, researchers should consider whether the meaning of the data is retained and whether any coding would be uninformative, or worse, misleading.

Applying the coding manual: training coders. To successfully code open-ended materials, researchers usually need to be trained to use the coding categories in the coding manual effectively. This training may happen in conjunction with the development of the coding manual or separately from its development (Adler et al., 2017); training coders can often reveal opportunities to refine the coding manual in a dynamic process of disagreements, confusions, discussions, and resolutions. Therefore, documentation of the training procedures can be helpful for detailing how the categories are defined conceptually. For example, we recommend selecting examples for the coding manual and the manuscript during the process of developing the manual and training coders (Adler et al., 2017).

When defining coding categories, it is useful to have a theory or other criteria to use as a benchmark for categories. One way this is done is to have a "gold-standard coder" who has content-matter expertise. With a gold-standard coder, the goal of training and agreement metrics is for coders to match the codes of the gold-standard coder (Bakeman & Quera, 2011; Gwet, 2014; Syed & Nelson, 2015), which provides the advantage of establishing agreement based on expertise rather than on matching alone. With matching alone as a

criterion for reliability, it is possible that matches between coders are based on a common misunderstanding of the theory or data, resulting in reliable but not valid coding. A gold-standard coder who has expertise helps to ensure that matches reflect valid coding.

Another option is to include a coder who is less familiar with the study materials. For example, when Campbell and colleagues (2023) de-identified transcripts of sexual-assault survivors, her team followed up by examining whether the de-identified transcripts were still valid representations of the participants' meaning. To do this, they asked a coder who had not been involved in the interviews to interpret the data. Including uninvolved coders can test whether data would be understandable to someone without prior knowledge of the study. Decisions about who is coding can therefore improve validity and clarity of the data, and documenting the characteristics of the coders can aid in evaluations of the findings.

Positionality: disclosing the researcher as an instrument. In addition to expertise or familiarity with the data, researchers' backgrounds and perspectives based on their identities may shape the research process and development of coding categories (Jamieson et al., 2023; Peterson, 2019; Steltenpohl, 2020). Therefore, some researchers include positionality statements to convey their potential biases, their status and privilege, and how they intend to remain aware of these factors (e.g., by having others review their coding efforts; Holmes, 2020; Jafar, 2018; Shaw, 2010). These statements can encourage reflexivity, enhance transparency about the research process, and amplify the voices of marginalized scholars (Elsherif et al., 2022; Manalili et al., 2023; Yang et al., 2022).

However, it is also important to consider that disclosing the authors' identities does not ensure impartiality (Savolainen et al., 2023). Furthermore, marginalized individuals could be targeted prejudicially because of their identity characteristics and therefore may not want to make their identities public knowledge (Darda et al., 2023; Oswald, 2024). This can be especially complex for people with intersecting dimensions of identity (e.g., race, neurodiversity, and gender), which are not favored by those in the dominant narrative (Sedgewick et al., 2021).[3] For these reasons, researchers should weigh the potential benefits of positionality statements against their possible drawbacks before deciding whether to add such a statement to a manuscript.

Establishing and reporting agreement

Reliability can be established in several ways,[4] including by consensus, but we focus here on interrater reliability. Interrater reliability is a formal metric of agreement between independent coders that quantifies the extent to which the coders assign the same codes to the same phenomena. Often, this involves two or more individuals independently coding a subset of the data (typically 10%–30%) to establish agreement.

Practice coding must be distinct from coding that is a test of reliability. For example, a practice phase that is particularly successful cannot be called a test phase post hoc without inflating many measures of agreement (see the section on kappa-hacking below). One way to draw a clear line between practice coding and the test of interrater reliability is for researchers to specify preestablished criteria that determine when coders are ready for the test phase of reliability. For example, the reliability test phase can be planned for when a criterion is met (e.g., when agreement is over 90% among all coders twice in a row) or on a specific date. Researchers may decide to include the criteria or other markers of the test phase in a preregistration. In addition, records should be kept to indicate which data were used for training and which were used for testing agreement because responses that are discussed in training should not be used in the test phase because this would inappropriately inflate the measure of agreement (Hallgren, 2012).

Certain features of the data can affect—either implicitly or explicitly—coders' decisions. To address this, in a process often referred to as "masking," researchers may decide to make certain information unavailable to the coders (e.g., responses to other questions that are not being coded, demographic information, experimental condition). Effectively, then, the coders' decisions are made irrespective of these features of the overall data set. However, there may be cases when a holistic understanding of the data requires consideration of identity, responses to related questions, or other information. Therefore, the decision to mask should be made based on the specific research project. It is important to report what was masked or not masked in the test phase of interrater reliability because this information affects replication and validity.

Calculating a metric of interrater reliability. There are several statistical measures for calculating interrater reliability. These include Cohen's κ, for pairs of raters (Cohen, 1960); Fleiss's κ, an adaptation of Cohen's κ for three or more raters (Fleiss, 1971); intraclass correlation coefficients (ICCs); and percentage agreement, among others (e.g., Siegel & Castellan's κ, Siegel & Castellan, 1988; Krippendorff's κ, Krippendorff, 1970). Whichever metric is chosen to report reliability, it should be justified in terms of its compatibility with the structure and distribution of the data and the coding.

Putting in place the cutoff for acceptable agreement before entering the testing phase, perhaps in a

preregistration, can also protect researchers from shifting their standards to suit the results of the coding process, especially because both Cohen's κ and ICCs lack clear guidelines for interpreting their magnitude.[5] Therefore, we suggest that researchers consider the goals and features of their research questions when determining an acceptable level of agreement. When faulty coding could create serious harm, higher levels of agreement may be needed (Conry-Murray & Silverstein, 2022).

If the first test phase fails to establish agreement, it may be necessary to attempt another test phase. The number of tests required to pass the established threshold should be disclosed because multiple tests can inflate many measures of agreement (see the section on kappa-hacking, below). To improve agreement without inappropriately inflating it, researchers should (a) develop a strong coding manual, (b) discuss and improve codes before the test round (Hemmler et al., 2022), and (c) select the most accurate coders before the test round. For a summary of methods for improving agreement that range from acceptable to fraudulent, see Table 2.

Resolving coding discrepancies. It can be helpful for coders to regularly discuss coding and examine discrepancies to prevent "drift," in which coders start to code differently when too much time has passed since training (Adler et al., 2017). However, even with continued discussion, it is possible for codes to include discrepancies between or among coders. Coding discrepancies in the test phase must be resolved before analyzing the data because typically only one set of data is analyzed. When there is a gold-standard coder, their codes should be prioritized because they have the most expertise on the topic. When there is not, discrepancies in coding are often decided by a process of discussion and consensus, or when data are continuous and codes are not widely discrepant, a mean may be taken. Another option is to have a third member of the research team code discrepancies, a method that works well if the team has several coders who have been shown to be reliable (Syed & Nelson, 2015). Again, these decisions should be disclosed and justified to allow audiences to evaluate their impact on the validity of the findings.

Coding remaining data. Coders typically divide the remaining coding among themselves to finish coding all data that were not part of the agreement test. It can be helpful to also disclose how this final coding was conducted. For example, because coding can be fatiguing, some researchers intersperse coding with other activities (Adler et al., 2017). When material is especially sensitive or potentially traumatizing (as in the case of coding sexual-assault survivors' stories, Campbell et al., 2023), it may be especially necessary to build in breaks.

Documentation of tacit coding practices can enhance the quality of the coding because decisions will likely be made more deliberately. It also makes it possible for readers to evaluate the validity of coding and potentially for future researchers to replicate the study. However, documentation and transparency alone are not enough to make coding valid. In the next section, we examine coding practices that can inflate agreement metrics, making coding appear more reliable and valid than it is.

Kappa-Hacking: Artificially Inflating Agreement Statistics

Some readers may be familiar with the concept of *p*-hacking in null hypothesis significance testing research (Simmons et al., 2011). A related concept, relevant to the present article, is kappa-hacking: the manipulation of agreement metrics to artificially inflate agreement until it reaches the required threshold, distorting the agreement metric. As we describe below, our investigation finds that kappa-hacking stems from two issues: (a) repeated measurements of agreement that are susceptible to high agreement by chance and (b) manipulations of categories to exploit easy-to-code but meaningless codes. Table 2 summarizes our findings about the ways these and other practices can result in either acceptable, ambiguous, problematic, or fraudulent reports of agreement. We examine these issues in detail below using simulations to investigate several ways that kappa can be manipulated to inflate perceived agreement.

Although the issues below apply to many metrics of agreement, we focus here on Cohen's κ, used with categorical data with pairs of coders, because it is one of the most common metrics of agreement (McHugh, 2012). However, Fleiss's κ, used with groups of coders coding categorical data, and ICCs, used with continuous data, are also common. All of these correct for how often raters agree by chance (Cohen, 1960; Hallgren, 2012), but they are all susceptible to inflation. Below, we discuss these other metrics in relation to our findings on kappa.

Kappa-hacking: a simulation study

To examine whether it is possible to falsely inflate kappa, we ran a simulation study using R (Version 4.2.2) on a Windows Server x64, using the packages *irr* (Gamer et al., 2012) and *dplyr* (Wickham et al., 2019), with 5,000 repetitions for each condition under which kappa can be calculated. Code used for this simulation is available on OSF (https://osf.io/du6gy/). These and the following simulation results are not intended to represent every case in which kappa could be calculated or fully show the degree to which kappa-hacking can influence interrater reliability. Rather, we use these simulation results

Table 2. Summary of Different Methods for Increasing Cohens' κ as a Measure of Interrater Agreement

Acceptability of practice		Description of practice
Acceptable practices	Training	When coders are in the practice coding phase, it is acceptable and encouraged to discuss responses, practice as much as needed, and calculate practice kappas to check for progress.
	Rejecting a coder before the agreement phase	If there are multiple coders being tested for reliability using kappa and some do not meet the threshold before the test phase, it is acceptable to use only the best coders in the test phase. This decision should be made before the test phase, which should be clearly communicated in coding reports.
	Rejecting a coder who does not match the "gold standard"	During the test phase, it is acceptable to retain only the coders who match the "gold-standard coder." Matching the gold-standard coder is not a reflection of random agreement as long as researchers are confident in the gold-standard coder's expertise and all kappas are reported.
	Adapting the coding categories based on the data	Researchers may want to test for specific categories that are part of the theory-based question being tested, but it is also appropriate to add new codes to reflect the nature of the data through a bottom-up process.
	Selecting the agreement statistic based on the number of coders	If only two coders will be used to code each observation (including the gold-standard coder), Cohen's κ should be used. If more than two coders will be used, Fleiss's κ should be used instead.
Ambiguous practices	Data peeking	In the agreement phase, checking kappas as the data are being coded does not inflate Cohen's κs because it is not sensitive to sample size. However, random matches are possible, and therefore, if data peeking is used with optional stopping and selective reporting, it can become misleading.
	Changing the threshold of acceptable agreement	Ideally, the threshold of acceptable agreement (often a κ of .70) should be set before the agreement phase. It should not be changed after viewing the results of the reliability testing.
	Changing the gold-standard coder after the test phase	The decision to use a gold-standard coder should not be changed as a method of increasing kappa without good conceptual or expertise-based reasons. This is because coders who match each other more than the gold-standard coder may have a common misconception that explains their agreement. However, if the decision is made to use a different gold-standard coder with more expertise, researchers can retest reliability using a new test phase.
	Collapsing the number of coding categories	Ideally, all coding categories will be determined before the test phase. It is problematic to add coding categories to increase kappa (e.g., adding "uncodable" or "missing" categories). However, it is acceptable to collapse categories if there is a conceptual or theory-based reason.
	Dummy-coding categories	It is possible to dummy-code variables that will be entered in a reliability analysis. It is acceptable to use presence or absence as long as all data from both presence and absence are from meaningful categories. When "absence" includes missing or uncodable data, it can inflate agreement.
Problematic practices	Reporting only the higshest Cohen's κ	It is misleading to report only the coders who reach the highest Cohen's κ in the test phase because the reported kappas may reflect coders who had high kappas based on chance.
	Repeatedly testing and reporting only the best Cohen's κ	Testing for Cohen's κ several times and reporting only the best kappa is unacceptable. This can inflate kappa because coders may have high agreement in one round by coincidence. If the test phase needs to be repeated, all kappas (including from the failed test phase) should be reported.

(continued)

Table 2. *(continued)*

Acceptability of practice		Description of practice
	Manipulating easily coded categories	It is problematic to introduce a coding category because it is easily codable and not because it is conceptually useful, because it can inflate Cohen's κ. For example, if missing data make up a significant portion of responses, counting missing codes as matches inflates agreement.
	Cherry-picking data to code	If cases are eliminated, the decision-making process should be reported, and the justification should not be to improve kappa.
Fraudulent practices	Fabricating data	Fabricated data are not valid, and they mislead readers. Reliability metrics should be calculated based on nonfraudulent data.
	Sharing codes among coders in independent coding	Interrater reliability that is reported as independent assumes coders do not share their coding during the test phase. Coders who have seen others' codes will not be able to make an independent selection.

to illustrate how these methods of kappa-hacking can change the results of interrater reliability in a study. Table 3 summarizes the simulations we conducted and results from each of them.

Does kappa account for chance agreement better than raw percentage? First, we note that kappa does an excellent job of accounting for truly random variation on average. We first simulated coding data from two coders for which there were two categories with equal base rates in the population (i.e., our simulated coders could code either Category A or Category B, where both Category A and Category B are equally likely). When the simulated coders were set to randomly assign categories, their proportion of agreement was close to .50, which is quite high considering the coders selected codes by chance. However, kappa takes random agreement into account, and indeed, the average kappa value in this simulation was

Table 3. Simulation Tests of Practices That Increase Cohen's κ

No.	Question	Test	Results	Kappa-hacking?
1	Does kappa account for chance agreement better than raw percent?	Two coders randomly assigning two categories with equal base rates	Percent agreement ≈ .50, κ ≈ 0. Kappa accounts for random agreement.	No
2	Can selective reporting of test phases inflate kappa?	Selectively reporting only the high pair in a group of coders	Fleiss's κ for mediocre coders ≈ .64, but when reporting only the best pair, Cohen's κ was > .70 24% to 52% of the time.	Yes
3	Can padding agreement with trivial categories inflate kappa?	Adding unneeded categories that are easy matches (e.g., missing data)	Kappas increased between .07 and .30 points when 25% of matches were because of unneeded easy matches.	Yes
4	Does consolidating categories inflate kappa?	Collapsing from three categories to two	Kappas increased .06 points when two categories were combined.	Yes
5	Does data peeking bias agreement scores?	Change in kappa between 50 and 300 cases when stopping at increments of 50	Kappas differed by .03 to .04 points	No
6	Does unevenly distributed data bias agreement scores?	Compared 50–50 base rates to 90–10	Kappa decreased by .40 points when base rates were skewed	No[a]

[a]Unevenly distributed data do not lead to kappa-hacking, but percentage agreement may be better when data are skewed.

close to 0, demonstrating the benefit of using kappa over percentage agreement.

Can selective reporting of test phases inflate kappa? Although Cohen's κ accounts for random agreement, agreement can still vary widely across tests, and kappa can be inappropriately inflated when only tests with high agreement are reported. For example, with four mediocre coders (we set the variance of the random error of mediocre coders to .35, compared with .30 for good coders), across 5,000 simulations, we found an average Fleiss's κ (an accurate estimate of group reliability) of .64. However, reporting only the best pair of coders from that group raised Cohen's κ to a falsely high .70 or above more than 30% of the time. With six coders, Fleiss's κ remained .64, but at least one pair of coders was able to report a Cohen's κ of .70 or above 52% of the time. Thus, mediocre coders can appear to be reliable when there is selective reporting about how many tests were conducted. If researchers did not identify the test phase and disclose the kappa scores for all coders and instead reported only the highest matches, then readers would not know that the coding was not reliable.

Using Fleiss's κ for multiple coders resolves some of the issues with multiple testing and selective reporting. However, when there is a gold-standard coder whose fidelity to the true data is higher, each individual coder is matched with a single expert, so Cohen's κ is appropriate. We therefore examined whether selective reporting of Cohen's κ is resolved when there is a gold-standard coder. We tested agreement with a gold-standard coder compared with four mediocre coders who each matched the gold-standard coder around 70% of the time. We used four categories and equal base rates of each category. The result was that the Fleiss's κ for all coders considered together was .57, but the chance of any coder getting Cohen's κ above .70 with the gold-standard coder was around 26%. Having a gold-standard coder therefore improves the accuracy when it represents true expertise but does not prevent kappa-hacking when selective reporting is used.

Our simulations indicate that when there are multiple coders, selective reporting can be misleading. To avoid cherry-picking kappas, authors can use practice rounds to select the best coders and use strategies such as preregistered criteria to decide ahead of time which round of coding will be the test round. When there are multiple rounds of tests (e.g., because of a failed test round), authors should report the agreement statistics for all test rounds.

When there are multiple coders, reporting Fleiss's κ resolved the issue of cherry-picking kappas. However, when there are pair-matches of coders, as with a gold-standard coder, Cohen's κ is needed. To avoid kappa inflation, researchers who have a gold-standard coder should take special care to report all kappas.

Can padding agreement with trivial categories inflate kappa? We also examined whether manipulations of the coding categories can falsely inflate Cohen's κ. Some coding categories, such as adding a category for missing data, make it easier to reach agreement; however, they may not add meaning to research questions. We tested whether kappa is inflated when easy-match categories that are not important to research questions or hypotheses are introduced.

In our 5,000 simulations, when codes were simulated randomly but included 25% easy matches (as might be the case when missing data are coded), it resulted in an inflation of Cohen's κ from the accurate measure of 0 to the inflated .30. With two mediocre coders and four categories, including 25% matched because of missing data, Cohen's κ was inflated from .37 to .55. With good coders and four categories, having 25% matches based on missing data inflated Cohen's κ from .76 to .83. We conclude that when missing data are not a meaningful coding category, using it as a coding category can falsely inflate kappa.

Matches based on missing data can arise unintentionally when "dummy-coding" categories, such as coding presence or absence of a single code. It is acceptable to use presence or absence of a single code as long as all data are from meaningful categories. When "absence" includes missing or other easy-to-code but not meaningful matches, it can inflate agreement between coders.

To address these concerns, one strategy is to remove missing or nonvalid responses from analyses first and then code the responses from meaningful categories. Authors should report which coding categories were used to calculate kappa. One rule of thumb is to calculate agreement metrics using only the categories that are related to the research questions.

Does consolidating categories inflate kappa? We further examined whether collapsing categories would inflate kappa. Indeed, with two mediocre coders and three categories (50% A, 30% B, 20% C and 200 observations), κ is about .65; however, combining Categories B and C results in an inflated κ of .71. We acknowledge that there may be legitimate methodological or theoretical reasons to collapse categories (e.g., when use of a category is very low and when categories can be meaningfully collapsed based on theory). However, collapsing categories solely to increase kappa is inappropriate. When categories are collapsed, they should be conceptually similar, and they should be the same categories that are then used for data analysis, reporting, and discussion.

Does data peeking bias agreement scores? We also examined whether adding more data affects Cohen's κ. In two simulations, we found only small changes in kappa resulting from adding more data. For example, with two mediocre coders, the average kappa across 5,000 simulations did not change substantially between 50 and 300 cases when stopping at increments of 50 cases (κ was between .530 and .534). In another trial, when stopping at the preselected number of 200, peeking at increments of 50 cases also did not substantially change kappa (average κ ranged between .639 and .642, depending on the sample size). Therefore, data peeking with Cohen's κ is generally unproblematic because it is relatively robust to changes in sample size. See Table 3.

However, we note that it is possible to achieve a high rate of matches in a single analysis by chance. Therefore, if data peeking is used in combination with optional stopping and selective reporting, it can still be a problem, even though data peeking is not a problem on average.

Does unevenly distributed data bias agreement scores? With just two categories, some have suggested that it can be challenging to get a sufficiently high Cohen's κ (McHugh, 2012; Syed & Nelson, 2015). We examined whether kappa provides useful information when there are only two categories of codes. We simulated data with two equally likely categories for which coders were reasonably good. Given that two categories make it possible to match by chance frequently, some reduction in agreement seems appropriate. Indeed, their proportion of agreement was .82, and their Cohen's κ was .67, averaged over 5,000 simulations.

However, it is more difficult to reach an acceptable kappa when the base rate prevalence is not equal. For example, when there is a 90% chance of one category and 10% chance of the other, our simulations showed that the proportion of agreement between two good coders remained around .82, but κ decreased to .42, on average, over the 5,000 simulations. Even when we used Fleiss's κ, the proportion of agreement was still substantially higher in every simulation condition. Therefore, we suggest that percentage or proportion agreement could be reported in addition to Cohen's κ when there are only two categories and especially when base rates are unequal (see also Xu & Lorber, 2014).

Other measures of interrater reliability

The ICC offers a means of reporting interrater reliability when data are continuous (i.e., ordinal or interval). ICC assesses reliability by measuring correlations within a class, such as repeated measures (Liljequist et al., 2019). Fleiss's κ also allows for more than two coders, but it is used with categorical data. Although similar in intent, these measures are based on slightly different information (e.g., kappa reflects degree of agreement, whereas the ICC can show consistency between raters or degree of agreement; Shrout & Fleiss, 1979).

The same general considerations that apply to the use of Cohen's κ for interrater reliability also are relevant when using the ICC, Fleiss's κ, and other metrics. In particular, when data that are not meaningful are included only for the purpose of improving agreement (e.g., including missing data as matches), it can inflate any measure of agreement. In addition, although most measures of agreement account for chance agreement over many tests, individual tests may occasionally be high because of chance. If agreement by any measure is reported selectively (e.g., by repeated testing with different samples or by selecting different constellations of coders), it can lead to artificially high metrics.

Therefore, we suggest that authors should report all relevant design considerations and analytic results to reassure the reader that no such manipulation has taken place. Just as with the interpretation of other metrics, such as effect sizes, prevailing discipline-specific standards also should be taken into account in interpreting the meaning of any interrater reliability values.

Transparent Practices: Recommendations for Journals and Authors

In the sections above, we discussed several practices that often go unreported in research with quantitatively coded open-ended data (e.g., reporting the reliability of identifying units, reporting all kappas in a test phase, reporting all categories tested for reliability). We cannot be sure whether these practices for rigorous coding development, training, and application are common because, as we found in our analysis of published articles, most articles provide very little information about how their coding was conducted. If researchers were more transparent about research practices and if journals supported this effort, it would help readers to assess whether the coding reflects valid and reliable categories.

However, the nature of open-ended data varies greatly across different types of studies (e.g., autobiographical reasoning in an interview study, observations of child behavior, gaze patterns from a visual-expectancy study). Accordingly, the challenges, benefits, and practicalities of implementing transparent practices will also vary greatly across studies (e.g., coding narrative data from autobiographical interviews might call for a more elaborate coding manual than coding infant eye-gaze patterns in an experimental study). We caution against unilaterally adopting the following recommendations without considering the context of each research study. Keeping

these caveats in mind, below, we discuss three areas in which more transparency could potentially increase trust in research with open-ended data.

Sharing materials

One way to improve transparency is to share materials on platforms such as OSF. We recommend sharing materials such as the coding manual, reliability coding data and script, and final coding data and scripts. Sharing the full coding manual can help to document how the researchers established coding categories. Sharing the de-identified data and script for interrater reliability can allow others to assess the statistical validity of the agreement metric. Sharing the statistical code or script documents how the data were used for hypothesis testing or exploration. In addition, sharing the rationale for the many tacit decisions can help make the choices better understood.

When privacy and copyright concerns allow, it can also be helpful to openly share the raw data being coded. Sharing private or identifiable data requires participants' consent. As appropriate, researchers may need to explicitly ask for consent to make data available on a public website for analyses beyond the current project.

Narrative data can be difficult to anonymize, and consent forms can address this by explaining how participants' data will be shared with raters and publicly, for example, in presentations or in a data repository (Adler et al., 2017; Campbell et al., 2023). When topics are especially sensitive, sharing with participants exactly how data will be de-identified can help them to decide what information is safe to share with researchers (Campbell et al., 2023).

Even with informed consent to share open-ended data, researchers should make a serious effort to de-identify data, especially when the safety of participants is at stake. Although direct identifiers, such as social security numbers, are clearly problematic, it is also possible to identify participants through voice or images or through combinations of seemingly innocuous variables, such as race, gender, and age. Furthermore, when some data come from focus groups or other interactions, it may be possible for "insiders," such as a domestic-violence perpetrator, to identify participants (see Campbell et al., 2023). Some researchers also offer participants the option to redact specific statements at the end of the session. Campbell et al. (2023) created a codebook of the types of information that might be identifiable using guiding questions to evaluate ambiguous information (e.g., "Who else would know this information?" "How would they know?" "What other records exist that contain the information?"). Coders attempted first to blur the information by altering specific words to make them more general. If that was not sufficient, coders used redaction.

When data need to be restricted but can otherwise be made available to people outside the research team, it can be shared via repositories that are not public, such as Databrary (https://nyu.databrary.org/), where access to sensitive or identifiable data requires access control. Researchers should also take special care regarding what laws apply regarding personal data (e.g., the Family Educational Rights and Privacy Act in the United States and General Data Protection Regulation in the European Union).

Usability of shared materials

Transparency includes usability. Too much information (e.g., providing copies of every iteration of the coding manual) or information that is hard to navigate may not be useful. We therefore recommend making information as user-friendly as possible (e.g., including a README file with data set name and structure, variable names and how they are coded, and any other helpful information for getting started).

When done ethically and carefully, transparent reporting can allow people who are not affiliated with the project to understand what is being claimed, coded, and reported, which allows for better assessments of validity and replication.

Preregistration

Given that it is possible to inappropriately inflate measures of agreement and otherwise hide bias in coding, there are advantages to prespecifying plans and hypotheses through preregistration or Registered Reports. Registered Reports have the additional advantage of allowing researchers to receive external peer feedback on their design and preregistration decisions (e.g., Chambers & Tzavella, 2022; Henderson & Chambers, 2022; Nosek et al., 2018). Prespecifying some of the process-level decisions can be helpful for avoiding inflated kappas and can help to transparently communicate which data-analysis decisions were made in advance.

We provide a template for the preregistration of coding open-ended data quantitatively (shared on OSF: https://osf.io/du6gy/). Some examples of information to include in the preregistration template are information about the plans for establishing agreement or interrater reliability (including information on practice and test phases), whether there will be a gold-standard coder, and what the threshold for agreement will be, among other things. However, we also acknowledge that each study has different features and goals and that authors

Justifications were coded by a gold-standard coder (the third author) who is an expert in social domain theory and has extensive coding experience. The gold-standard coder and three trained research assistants first established the units of analysis, with agreement established in the test phase of between 91-96%, above the 90% agreement threshold established in our pre-registration. Each unit received only one code. The coding categories that were included in the test phase are available in SOM. Inter-rater reliability for coding of the content of the units of analysis was tested by comparing each research assistant's independent coding to the gold standard coding in a single test phase, which included 20% of interviews (40 participants' interviews). No interviews in the test phase had been previously discussed, but all coders had access to the full interview transcript. Two of the three research assistants matched with the gold-standard coder at Cohen's kappa = .76 and .81, over the pre-registered threshold of .70. Those two research assistants coded the remaining 80% of interviews. Codes from the test phase were retained from the gold-standard coder.

Fig. 3. A sample of the proposed format for reporting interrater reliability.

should decide what is useful to preregister based on their specific study.

Reporting coding in the manuscript

Finally, we suggest that reporting in articles, even with limited space, could be more transparent than the current practices we found. Specifically, we recommend that authors regularly report (a) the method of selecting materials and when they are ambiguous, the method of establishing units; (b) the coding categories used in interrater-reliability metrics and analysis of research questions, their definition, and inclusion and exclusion criteria; (c) the qualifications of the coders, including whether there was a gold-standard coder and potentially, positionality; (d) the results of all agreement tests from the test phase(s) and how they were conducted (e.g., masked, independent); and (e) how the final codes were determined. For an illustration of such a report, see Figure 3.

In Table 4, we offer more details about what information we recommend should be disclosed in the article and supplemental materials, aiming to distill the considerations we covered throughout this article into specific strategies for authors that supplement our preregistration template (see SOM, https://osf.io/du6gy/). We also offer guidance to help editors and reviewers through the process of evaluating this work described in this article for increasing the validity of quantitative coding of open-ended data. The proposed guidelines are of a global nature, and we encourage editors, reviewers, and authors to tailor them for the specific needs and challenges in their relevant research communities. Because each study has different goals and different researchers have different resource constraints, each research team should consider how their practices aid reproducibility and allow readers to make accurate judgments about the validity of the research. We do not believe that the questions in Table 4 should be criteria that are applied to every study. Instead, researchers and editors should consider the specific features of each study and the context in which implications may apply. Therefore, the questions from Table 4 should be adapted to the specific study.

Conclusion

The rich nature of open-ended data means that it can provide deep insights into participants' thinking and behavior. Indeed, we are glad to see that open-ended data are commonly used in the journals we assessed. However, a key challenge for doing valid research is to represent open-ended data accurately and credibly. When topics are sensitive or have implications for policy or could provide the basis of future research, it is especially important to convey the meaning credibly. Questionable research practices in the quantitative coding of open-ended data—many of which we have seen in practice—can undermine scientific progress and harm those affected by the research. Unfortunately, with current standards, these questionable practices remain hidden from research reports.

Increased transparency allows other researchers to assess the fidelity of the interpretation of data. With better documentation, researchers could have the information they need to evaluate whether data accurately represent participants' thoughts and behaviors. It also provides important information to researchers who want to reproduce or generalize their findings to other contexts.

Increasing transparency about coding practices can also help establish norms for quality coding practices to ensure that valid practices are adopted by the broader research community. Our analysis of current practices in peer-reviewed journals suggests that transparent practices are rare. Academic journals could increase their support for transparency by encouraging more thorough reporting of important information about the coding processes. Authors, reviewers, and editors alike can push

Table 4. Questions to Consider Across Different Stages of Coding Open-Ended Data

Research stage	Questions for authors	Questions for editors and reviewers
Design and planning	What is the rationale for choosing quantitatively coding open-ended data over other methods explained? How will material be selected for coding? What units will be coded? How will they be identified reliably? How will agreement be established so that the coding is replicable and valid? What benchmarks (e.g., Cohen's κ) will be used for agreement? Will the plan for establishing agreement be prespecified in a preregistration or Registered Report? Where will documentation about the coding process be kept, and who will be responsible for producing it?	Was the coding of open-ended data preregistered? Were the selection of material, units, and agreement metrics reasonable, and did they match the plans in the preregistration (or were deviations explained)? If the study was not preregistered, were the decisions regarding the selection of material, units, and agreement metrics justified?
Coding-manual development and training	What coding categories will be used? Do the categories reflect the full range of responses or theory-based concepts? Will units each get one code or multiple codes? How will missing or incomplete data be handled? How will training be distinguished from the reliability test phase? Will there be a "gold-standard coder"? Will you include relevant positionality statements?	If full materials were not coded, was the selection of material justified? Are the coding categories available to examine, with examples and inclusion and exclusion criteria? How many codes did each unit get? How was missing or incomplete data handled?
Agreement	Does reliability need to be reported for the selection of units? How will practice coding be distinguished from the test phase? Which statistic will be most appropriate to assess interrater reliability? What criteria will be used to determine if coders have a sufficient level of agreement? How will discrepancies be handled?	Was interrater reliability established for the unit of analysis (if units were ambiguous)? Was agreement tested with a gold-standard coder? Was the test phase for agreement distinguished from practice coding? Was the chosen agreement statistic appropriate to assess interrater reliability? How many tests were done to establish reliability? Was the interrater reliability test phase masked, with new material that is representative of the data set? Are the practice and especially the test-phase codes data available? Is the statistical code for reliability testing available? What criteria was used to determine if coders have a sufficient level of agreement? How were discrepancies handled?

for systemic improvement in psychology by encouraging students, trainees, colleagues, and journals to consider the standards we have recommended. We, as researchers, can fulfill our responsibility to participants and others affected by research only if we strive to conduct valid research.

Transparency

Action Editor: David A. Sbarra
Editor: David A. Sbarra
Author Contribution

Clare Conry-Murray: Conceptualization; Formal analysis; Methodology; Project administration; Supervision; Writing – original draft; Writing – review & editing.
Tal Waltzer: Conceptualization; Data curation; Formal analysis; Methodology; Project administration; Visualization; Writing – original draft; Writing – review & editing.
Fiona C. DeBernardi: Data curation; Formal analysis; Methodology; Visualization; Writing – original draft; Writing – review & editing.
Jessica L. Fossum: Data curation; Formal analysis; Methodology; Software; Visualization; Writing – original draft; Writing – review & editing.
Simona Haasova: Conceptualization; Writing – original draft; Writing – review & editing.
Michael S. Matthews: Conceptualization; Writing – original draft; Writing – review & editing.
Aoife O'Mahony: Conceptualization; Writing – original draft; Writing – review & editing.

David Moreau: Conceptualization; Writing – original draft; Writing – review & editing.
Myriam A. Baum: Conceptualization; Writing – original draft; Writing – review & editing.
Veli-Matti Karhulahti: Conceptualization; Writing – original draft; Writing – review & editing.
Randy J. McCarthy: Conceptualization; Writing – original draft; Writing – review & editing.
Helena M. Paterson: Conceptualization; Visualization; Writing – original draft; Writing – review & editing.
Kara McSweeney: Conceptualization; Writing – original draft; Writing – review & editing.
Mahmoud M. Elsherif: Conceptualization; Visualization; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was partially supported by a National Science Foundation grant awarded to author T. Waltzer (NSF SPRF-FR# 2104610).

Open Practices

This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID IDs

Clare Conry-Murray https://orcid.org/0000-0003-0245-1289
Michael S. Matthews https://orcid.org/0000-0003-1695-2498
David Moreau https://orcid.org/0000-0002-1957-1941
Helena M. Paterson https://orcid.org/0000-0001-7715-5973
Mahmoud M. Elsherif https://orcid.org/0000-0002-0540-3998

Acknowledgments

This article began as an unconference at the Society for the Improvement of Psychological Science (2021). We thank Jan B. Vornhagen, who worked with T. Waltzer to create a demonstration of the process for developing a coding manual, which is shared on the OSF page for this article (https://osf.io/du6gy/). We also thank Annayah Prosser and Richard Klein for their contributions during the early stages of conceptualizing and writing this article. We also thank Kyra Larson, Maryam (Aria) Arianezhad, Audun Dahl, and members of the Developmental Moral Psychology Lab at UC Santa Cruz for their inputs on earlier versions of this article.

Notes

1. Some other articles contained brain- or audio-imaging data (8%), which could potentially be considered open-ended, but because the standard approaches for handling such data are different from the aim of the present research, we decided not to examine them further.
2. Although many types of data were unique to each study, some types of data did come up repeatedly. For example, text-box or interview responses, video recording of behaviors, and eye tracking were common. However, the nature of open-ended data varied across the different journals that we analyzed. See SOM for more details on types of open-ended data in each journal.
3. We refer interested readers elsewhere for further discussion on the role of positionality in research (e.g., Boghossian, 2006; Elsherif et al., 2022; Karhulahti, 2024; Puthillam et al., 2024; Savolainen et al., 2023).
4. For further details about the different methods of establishing agreement, see Bakeman and Quera (2011), Syed and Nelson (2015), and Tinsley and Weiss (1975).
5. For example, guidelines for kappa are only briefly mentioned in the Standards for Educational & Psychological Testing (American Psychological Association, 2019). For kappa, many journals use .70 as a cutoff for acceptable reliability, whereas other sources suggest > .80 or > .90 as appropriate criteria (McHugh, 2012; for tabulated values commonly accepted for different measures of agreement, also see Watts and Finkenstaedt-Quinn, 2021, p. 568). Excellent reliability is indicated by ICC values whose 95% confidence-interval values fall into the range above 0.9, although in some disciplines, a range above 0.75 may also be considered good to excellent (see Bryer, 2023, Table 2).

References

Adler, J. M., Dunlop, W. L., Fivush, R., Lilgendahl, J. P., Lodi-Smith, J., McAdams, D. P., McLean, K. C., Pasupathi, M., & Syed, M. (2017). Research methods for studying narrative identity: A primer. *Social Psychological and Personality Science*, *8*(5), 519–527. https://doi.org/10.1177/1948550617698202

Adu, P. (2019). *A step-by-step guide to qualitative data coding*. Routledge.

Aguinis, H., & Solarino, A. M. (2019). Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal*, *40*, 1291–1315. https://doi.org/10.1002/smj.3015

American Psychological Association. (2019). *The standards for educational and psychological testing*. https://www.apa.org/science/programs/testing/standards

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9781139017343

Boghossian, P. (2006). *Fear of knowledge: Against relativism and constructivism*. Oxford University Press.

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. Sage.

Bryer, J. (2023). *Relationship between intraclass correlation (ICC) and percent agreement*. https://irrsim.bryer.org/articles/IRRsim.html

Campbell, R., Javorka, M., Engleton, J., Fishwick, K., Gregory, K., & Goodman-Williams, R. (2023). Open-science guidance for qualitative research: An empirically validated approach for de-identifying sensitive narrative data. *Advances in*

Methods and Practices in Psychological Science, *6*(4). https://doi.org/10.1177/25152459231205832

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Collins, H. M. (2001). Tacit knowledge, trust and the Q of sapphire. *Social Studies of Science*, *31*(1), 71–85. https://doi.org/10.1177%2F030631201031001004

Conry-Murray, C., & Silverstein, P. (2022). *The role of values in psychological science: Examining identity-based inclusivity*. PsyArXiv. https://doi.org/10.31234/osf.io/cskg2

Conry-Murray, C., & Turiel, E. (2012). Jimmy's baby doll and Jenny's truck: Young children's reasoning about gender norms. *Child Development*, *83*(1), 146–158. https://doi.org/10.1111/j.1467-8624.2011.01696.x

Darda, K. M., Conry-Murray, C., Schmidt, K., Elsherif, M. M., Peverill, M., Yoneda, T., Lawson, K. M., DiGregory, E., Moreau, D., Shorter, G. W., & Gernsbacher, M. (2023). *Promoting civility in formal and informal open science contexts*. PsyArXiv. https://doi.org/10.31234/osf.io/rfkyu

DeCuir-Gunby, J. T., Marshall, P. L., & McCulloch, A. W. (2011). Developing and using a codebook for the analysis of interview data: An example from a professional development research project. *Field Methods*, *23*(2), 136–155. https://doi.org/10.1177%2F1525822X10388468

DeKeseredy, W. S., Stoneberg, D. M., Nolan, J., & Lory, G. L. (2020). Improving the quality of survey data on college campus woman abuse: The contribution of a supplementary open-ended question. *Violence Against Women*, *27*(12–13), 2477–2490. https://doi.org/10.1177/1077801220975496

Elman, C., & Kapiszewski, D. (2014). Data access and research transparency in the qualitative tradition. *Political Science & Politics*, *47*(1), 43–47. https://doi.org/10.1017/S1049096513001777

Elsherif, M. M., Middleton, S. L., Phan, J. M., Azevedo, F., Iley, B. J., Grose-Hodge, M., Tyler, S., Kapp, S. K., Gourdon-Kanhukamwe, A., Grafton-Clarke, D., Yeung, S. K., Shaw, J. J., Hartmann, H., & Dokovova, M. (2022). *Bridging neurodiversity and open scholarship: How shared values can guide best practices for research integrity, social justice, and principled education*. MetaArXiv. https://doi.org/10.31222/osf.io/k7a9p

Fine, G. A., & Elsbach, K. D. (2000). Ethnography and experiment in social psychological theory building: Tactics for integrating qualitative field data with quantitative lab data. *Journal of Experimental Social Psychology*, *36*(1), 51–76. https://doi.org/10.1016/j.obhdp.2020.10.015

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). Package 'irr'. In *Various coefficients of interrater reliability and agreement*, *22*, 1–32. https://cran.r-project.org/web/packages/irr/irr.pdf

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *The American Scientist*, *102*(6). https://doi.org/10.1511/2014.111.460

Grimm, P. (2010). Social desirability bias. In *Wiley international encyclopedia of marketing*. John Wiley. https://doi.org/10.1002/9781444316568.wiem02057

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hales, A. H., Wesselmann, E. D., & Hilgard, J. (2019). Improving psychological science through transparency and openness: An overview. *Perspectives on Behavior Science*, *42*, 13–31. https://doi.org/10.1007/s40614-018-00186-8

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%2Ftqmp.08.1.p023

Hemmler, V. L., Kenney, A. W., Langley, S. D., Callahan, C. M., Gubbins, E. J., & Holder, S. (2022). Beyond a coefficient: An interactive process for achieving inter-rater consistency in qualitative coding. *Qualitative Research*, *22*(2), 194–219. https://doi.org/10.1177/1468794120976072

Henderson, E. L., & Chambers, C. D. (2022). Ten simple rules for writing a Registered Report. *PLOS Computational Biology*, *18*(10), Article e1010571. https://doi.org/10.1371/journal.pcbi.1010571

Holmes, A. G. D. (2020). Researcher positionality - A consideration of its influence and place in qualitative research - A new researcher guide. *International Journal of Education*, *8*(4), 1–10. https://doi.org/10.34293/education.v8i4.3232

Howitt, D., & Cramer, D. (2011). *Introduction to research methods in psychology* (3rd ed.). Prentice Hall.

Hruschka, D. J., Schwartz, D., St John, D. C., Picone-Decaro, E., Jenkins, R. A., & Carey, J. W. (2004). Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field Methods*, *16*(3), 307–331. https://doi.org/10.1177/1525822X04266540

Jackson, K. M., & Trochim, W. M. (2002). Concept mapping as an alternative approach for the analysis of open-ended survey responses. *Organizational Research Methods*, *5*(4), 307–336. https://doi.org/10.1177/109442802237114

Jafar, A. J. N. (2018). What is positionality and should it be expressed in quantitative studies? *Emergency Medicine Journal*, *35*(5), 323–324. https://doi.org/10.1136/emermed-2017-207158

Jamieson, M. K., Govaart, G. H., & Pownall, M. (2023). Reflexivity in quantitative research: A rationale and beginner's guide. *Social and Personality Psychology Compass*, *17*(4), Article e12735. https://doi.org/10.1111/spc3.12735

Karhulahti, V.-M. (2022). Reasons for qualitative psychologists to share human data. *British Journal of Social Psychology*, *62*(4), 1621–1634. https://doi.org/10.1111/bjso.12573

Karhulahti, V. M. (2024). Positionality statements in science. *Open Research Europe*, *4*, 62. https://doi.org/10.12688/openreseurope.17058.1

Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE Guide No. 131. *Medical Teacher*, *42*(8), 846–854. https://doi.org/10.1080/0142159X.2020.1755030

Krippendorff, K. (1970). Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70. https://doi.org/10.1177/001316447003000105

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation–A discussion and demonstration of basic features. *PLOS ONE*, *14*(7). https://doi.org/10.1371/journal.pone.0219854

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

Linneberg, M. S., & Korsgaard, S. (2019). Coding qualitative data: A synthesis guiding the novice. *Qualitative Research Journal*, *19*(3), 259–270. https://doi.org/10.1108/QRJ-12-2018-0012

López-Nicolás, R., López-López, J. A., & Rubio-Aparicio, M. (2022). A meta-review of transparency and reproducibility-related reporting practices in published meta-analyses on clinical psychological interventions (2000–2020). *Behavioral Research Methods*, *54*, 334–349. https://doi.org/10.3758/s13428-021-01644-z

MacQueen, K. M., McLellan, E., Kay, K., & Milstein, B. (1998). Codebook development for team-based qualitative analysis. *Cam Journal*, *10*(2), 31–36. https://doi.org/10.1177/1525822X980100020301

Manalili, M. A. R., Pearson, A., Sulik, J., Creechan, L., Elsherif, M., Murkumbi, I., Azevedo, F., Bonnen, K. L., Kim, J. S., Kording, K., Lee, J. J., Obscura, M., Kapp, S. K., Röer, J. P., & Morstead, T. (2023). From puzzle to progress: How engaging with neurodiversity can improve cognitive science. *Cognitive Science*, *47*(2), Article e13255. https://doi.org/10.1111/cogs.13255

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

Natow, R. S. (2022). Policy actors' perceptions of qualitative research in policy making: The case of higher education rulemaking in the United States. *Evidence & Policy*, *18*(1), 109–126.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, USA*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Oswald, F. (2024). Positionality statements should not force us to 'out' ourselves. *Nature Human Behaviour*, *8*, Article 185. https://doi.org/10.1038/s41562-023-01812-5

Peterson, J. S. (2019). Presenting a qualitative study: A reviewer's perspective. *Gifted Child Quarterly*, *63*(3), 147–158. https://doi.org/10.1177/0016986219844789

Pownall, M., Talbot, C. V., Kilby, L., & Branney, P. (2023). Opportunities, challenges and tensions: Open science through a lens of qualitative social psychology. *British Journal of Social Psychology*, *62*(4), 1581–1589. https://doi.org/10.1111/bjso.12628

Puthillam, A., Montilla Doble, L. J., Delos Santos, J. J. I., Elsherif, M. M., Steltenpohl, C. N., Moreau, D., Pownall, M., Silverstein, P., Anand-Vembar, S., & Kapoor, H. (2024). Guidelines to improve internationalization in the psychological sciences. *Social and Personality Psychology Compass*, *18*(1), Article e12847. https://doi.org/10.1111/spc3.12847

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. In A. Ferligoj & A. Mrvar (Eds.), *Developments in applied statistics* (pp. 159–177). University of Ljubljana. http://mrvar.fdv.uni-lj.si/pub/mz/mz19/reja.pdf

Ruona, W. E. (2005). Analyzing qualitative data. In R. E. Swanson & E. F. Holton, III (Eds.), *Research in organizations: Foundations and methods of inquiry* (pp. 233–264). Berrett-Koehler. https://doi.org/10.1207/s15430421tip3903_5

Ryu, S. (2020). The role of mixed methods in conducting design-based research. *Educational Psychologist*, *55*(4), 232–243. https://doi.org/10.1080/00461520.2020.1794871

Saldaña, J. (2016). *The coding manual for qualitative researchers* (3rd ed.). Sage.

Savolainen, J., Casey, P. J., McBrayer, J. P., & Schwerdtle, P. N. (2023). Positionality and its problems: Questioning the value of reflexivity statements in research. *Perspectives on Psychological Science*, *18*(6), 1331–1338. https://doi.org/10.1177/17456916221144988

Sedgewick, F., Hull, L., & Ellis, H. (2021). *Autism and masking: How and why people do it, and the impact it can have*. Jessica Kingsley Publishers.

Shaw, R. (2010). Embedding reflexivity within experiential qualitative psychology. *Qualitative Research in Psychology*, *7*(3), 233–243. https://doi.org/10.1080/14780880802699092

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). McGraw-Hill Book Company.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Singer, E., & Couper, M. P. (2017). Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys. *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology*, *11*(2), 115–134. https://doi.org/10.12758/mda.2017.01

Smyth, J. D. (2016). Designing questions and questionnaires. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 218–235). SAGE Publications Ltd. https://doi.org/10.4135/9781473957893

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour*, *5*(8), 990–997.

Steltenpohl, C. N. (2020, August 20). *Is science objective?* [Blog post]. https://cnsyoung.com/is-science-objective/

Steltenpohl, C. N., Lustick, H., Meyer, M. S., Lee, L. E., Stegenga, S. M., Reyes, L. S., & Renbarger, R. (2023). Rethinking transparency and rigor from a qualitative open science

perspective. *Journal of Trial & Error*, *4*(1). https://doi.org/10.36850/mr7

Sussman, R. (2016). Observational methods: The first step in science. In R. Gifford (Eds.), *Research methods for environmental psychology* (pp. 9–27). John Wiley & Sons.

Syed, M., & Nelson, S. C. (2015). Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood*, *3*(6), 375–387. https://doi.org/10.1177/2167696815587648

Tenney, E. R., Costa, E., Allard, A., & Vazire, S. (2021). Open science and reform practices in organizational behavior research over time (2011 to 2019). *Organizational Behavior and Human Decision Processes*, *162*, 218–223. https://doi.org/10.1177/1525822X04266540

Tinsley, H. E., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*(4), 358–376. https://doi.org/10.1037/h0076640

van den Akker, O. R., van Assen, M. A., Bakker, M., Elsherif, M., Wong, T. K., & Wicherts, J. M. (2024). Preregistration in practice: A comparison of preregistered and non-preregistered studies in psychology. *Behavior Research Methods*, *56*, 5424–5433. https://doi.org/10.3758/s13428-023-02277-0

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, *31*(2), 162–168. https://doi.org/10.1177/09637214211067779

Waltzer, T., DeBernardi, F. C., & Dahl, A. (2023). Student and teacher views on cheating in high school: Perceptions, evaluations, and decisions. *Journal of Research on Adolescence*, *33*(1), 108–126. https://doi.org/10.1111/jora.12784

Waltzer, T., Samuelson, A., & Dahl, A. (2022). Students' reasoning about whether to report when others cheat: Conflict, confusion, and consequences. *Journal of Academic Ethics*, *20*, 265–287. https://doi.org/10.1007/s10805-021-09414-4

Watts, F. M., & Finkenstaedt-Quinn, S. A. (2021). The current state of methods for establishing reliability in qualitative chemistry education research articles. *Chemistry Education Research and Practice*, *22*(3), 565–578. https://doi.org/10.1039/D1RP00007A

Weston, C., Gandell, T., Beauchamp, J., McAlpine, L., Wiseman, C., & Beauchamp, C. (2001). Analyzing interview data: The development and evolution of a coding system. *Qualitative Sociology*, *24*(3), 381–400. https://doi.org/10.1023/A:1010690908200

Whittemore, R., Chase, S. K., & Mandle, C. L. (2001). Validity in qualitative research. *Qualitative Health Research*, *11*(4), 522–537. https://doi.org/10.1177/104973201129119299

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, Article 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H., François, R., Henry, L., Müller, K., & Wickham, M. H. (2019). *Package 'dplyr'. A grammar of data manipulation* (R Package Version 8). https://cloud.r-project.org/web/packages/dplyr/index.html

Wilson, D. B. (2009). Systematic coding. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (Vol. 2, pp. 159–176). Russell Sage Foundation.

Xu, S., & Lorber, M. F. (2014). Interrater agreement statistics with skewed data: Evaluation of alternatives to Cohen's kappa. *Journal of Consulting and Clinical Psychology*, *82*(6), 1219–1227. https://doi.org/10.1037/a0037489

Yang, Y., Tian, T. Y., Woodruff, T. K., Jones, B. F., & Uzzi, B. (2022). Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences, USA*, *119*(36), Article e2200841119. https://doi.org/10.1073/pnas.2200841119