Article

International Journal of Qualitative Methods
Volume 19: 1–13

DOI: 10.1177/1609406919899220
journals.sagepub.com/home/ijq
SAGE

Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines

Cliodhna O'Connor[1] and Helene Joffe[2]

Abstract

Evaluating the intercoder reliability (ICR) of a coding frame is frequently recommended as good practice in qualitative analysis. ICR is a somewhat controversial topic in the qualitative research community, with some arguing that it is an inappropriate or unnecessary step within the goals of qualitative analysis. Yet ICR assessment can yield numerous benefits for qualitative studies, which include improving the systematicity, communicability, and transparency of the coding process; promoting reflexivity and dialogue within research teams; and helping convince diverse audiences of the trustworthiness of the analysis. Few guidelines exist to help researchers negotiate the assessment of ICR in qualitative analysis. The current article explains what ICR is, reviews common arguments for and against its incorporation in qualitative analysis and offers guidance on the practical elements of performing an ICR assessment.

Keywords

intercoder reliability, interrater reliability, qualitative analysis, interviews, coding

Introduction

The power of qualitative research in shedding light on questions concerning experience and motivation is increasingly recognized by research funders and policy-makers. This growing popularity brings demand for specific, step-by-step guidelines on implementing the various stages of qualitative analysis. Several practical how-to guides have been published to date (e.g., Attride-Stirling, 2001; Braun & Clarke, 2006; Charmaz, 2006; Joffe, 2012; Roberts et al., 2019; Smith et al., 1999). Such resources enhance the accessibility and consistency of qualitative research and provide valuable teaching aids. However, few guidelines exist to help researchers navigate the assessment of intercoder reliability (ICR) in qualitative analysis. The current article seeks to fill this gap. It explains what ICR is, reviews common arguments for and against its incorporation in qualitative analyses, and offers guidance on the practical elements of performing ICR assessment.

The recommendations offered are based on a thorough review of the literature on ICR, as well as the authors' own research experience. The authors of this article are social scientists who conduct qualitative research to explore lay thinking, feeling, and experience concerning a range of topics including climate change (Smith & Joffe, 2009, 2013), emerging infectious diseases (Joffe, 1999; Joffe et al., 2011), neuroscience (O'Connor & Joffe, 2013, 2014a, 2014b, 2015; O'Connor et al., 2012), earthquakes (Joffe et al., 2013, 2018), sexuality (Lavie-Ajayi & Joffe, 2009; O'Connor, 2017), cities (Joffe & Smith, 2016), economics (O'Connor, 2012), and mental illness (O'Connor et al., 2018; O'Connor & McNicholas, 2019). Although the authors' primary disciplinary affiliation is social psychology, they have conducted qualitative research projects in many interdisciplinary contexts, involving collaborations with sociologists, anthropologists, engineers, psychiatrists and lawyers, among others. The current article is designed to be relevant to those working with qualitative techniques across the social sciences. It will be particularly useful to those new to ICR, though experienced researchers will also benefit from its review of the variety of perspectives on and practical approaches to ICR assessment.

In the authors' own research, data collection methods of choice have usually been in-depth interviews (often using Joffe

[1] School of Psychology, University College Dublin, Belfield, Dublin, Ireland
[2] Division of Psychology and Language Sciences, University College London, London, United Kingdom

Corresponding Author:
Cliodhna O'Connor, School of Psychology, University College Dublin, Belfield, Dublin 4, Ireland.
Email: cliodhna.oconnor1@ucd.ie

and Elsey's [2014] free association Grid Elaboration Method) and media analysis of both text and imagery (e.g. O'Connor & Joffe, 2014a; Smith & Joffe, 2009). Many of the examples offered in this article have these forms of data in mind; in particular in-depth interviews, since these pose particular challenges in ICR assessment and have received little specific attention in the ICR literature (Campbell et al., 2013). However, the ICR issues reviewed are broadly relevant to all forms of qualitative data, which include focus groups, free-text survey responses, diaries, and other written documents.

This article is equally broadly based regarding the analytic approach selected. In principle, ICR could be incorporated into any qualitative analysis that involves coding text or images. In practice, ICR is more popular within some analytic traditions than others. For instance, practitioners of content analysis often present ICR as a fundamental imperative of that method (Lombard et al., 2002; Neuendorf, 2002). In contrast, some grounded theory researchers see ICR as inappropriate due to the recursive, incremental nature of grounded theory's analytic process (Oktay, 2012). The question of ICR's epistemological compatibility with qualitative paradigms is discussed below. The authors of this article have most often applied ICR within thematic and content analysis and found it improved the quality, transparency, and reception of analyses.

What Is ICR?

ICR is a numerical measure of the agreement between different coders regarding how the same data should be coded. ICR is sometimes conflated with inter*rater* reliability (IRR), and the two terms are often used interchangeably. However, technically IRR refers to cases where data are rated on some ordinal or interval scale (e.g., the intensity of an emotion), whereas ICR is appropriate when categorizing data at a nominal level (e.g., the presence or absence of an emotion). Most qualitative analyses involve the latter analytic approach.

ICR should also be differentiated from *intra*coder reliability. This refers to consistency in how the same person codes data at multiple time points. That is, if the same person returns to the data at another time, will they code it consistently? Evaluating intracoder reliability may prove a useful exercise in promoting researcher reflexivity (Joffe & Yardley, 2003). However, it is not particularly common in qualitative research. When "reliability" is discussed, it usually refers to the intercoder level.

A further terminological distinction is between ICR and intercoder *consistency*. Many qualitative research teams include an element of comparison between individual team members' impressions of the data, but may refrain from quantifying the degree of consensus. For example, Thomas and Harden's (2008) approach of thematic synthesis suggests that independent researcher identification of themes could be followed by group discussion of overlaps and divergences. The term "reliability" implies that researchers have gone beyond this to formally compute a measure of intercoder agreement. The current article primarily focuses on quantified measures of reliability, since the existing literature offers little practical guidance on performing and interpreting ICR assessment. However, many of the issues discussed may also be relevant to researchers applying less structured evaluations of intercoder consistency.

ICR and the Coding Process

In the analysis phase of qualitative research, the social scientist must introduce a more conceptual understanding of the data (Gaskell, 2000). In most qualitative analyses, this involves the development of a *coding frame* that captures the analytically significant features of the data. The coding frame is typically a list of codes, which may be organized according to higher-order code categories, accompanied by code definitions and example data segments. The coding frame constitutes the analytic instrument with which the raw data is reduced, classified, and synthesized into a more conceptual framework (Gaskell, 2000). Once developed, the coding frame is applied systematically to the data. This means the data are segmented into data units and each data unit is labeled with codes that index its analytically relevant content. ICR can be calculated in the coding phase of qualitative analysis to assess the robustness of the coding frame and its application.

It is important to note that coding is just one stage in qualitative analysis. Codes can be conceptualized as the basic "building blocks" with which the structure of the analysis is constructed. After coding is completed, depending on the analytic approach used, codes are usually clustered into themes or narratives that are interpreted according to relevant theory. It is generally accepted that different analysts, with different theoretical commitments, will organize codes into themes in different ways (Armstrong et al., 1997). As long as researchers are transparent about their rationale for the thematic structure developed, this is not problematic; indeed, this level of interpretative flexibility is the *raison d'être* of qualitative research. The logic of applying ICR to the earlier coding phase is that coding is the first place where the analysis begins to move beyond the raw data into a more abstract conceptual framework. Haphazard or inappropriate coding at this stage fundamentally compromises the analysis' claims to offer a faithful and trustworthy characterization of the data. Qualitative researchers have proposed numerous different steps to substantiate the credibility of the coding process (Bauer et al., 2000; O'Brien et al., 2014; Popay et al., 1998; Seale & Silverman, 1997; Yardley, 2000). One among these is ICR: by increasing the consistency and transparency of the coding process, ICR can help provide confidence that specific efforts were made to ensure the final analytic framework represents a credible account of the data.

Current Practice Regarding ICR

Consideration of ICR is relatively common, although by no means ubiquitous, in qualitative research. A 2018 search for the key words "qualitative" and "intercoder reliability" or "inter-coder reliability" yielded over 1,000 results on Scopus,

and over 16,000 on Google Scholar. More specific information regarding the prevalence of ICR comes from the content analysis literature.[1] Lombard et al. (2002) report that of content analysis articles published in the mass communications literature between 1994 and 1998, 69% mentioned ICR. However, many reports of ICR calculations were vague and/or used inappropriate methods. An earlier audit of content analysis articles in consumer behavior and marketing journals between 1978 and 1989 found 48% used independent judges, but 31% reported no reliability coefficient and the method of calculating reliability was unclear in an additional 19% (Kolbe & Burnett, 1991). A more recent analysis of content analysis articles in two communications journals also found high levels of incomplete information and inappropriate testing and reporting practices (Feng, 2014). It should be noted that these studies relate only to the content analysis literature. Content analysis is the analytic tradition with the highest affinity for ICR; for instance, Neuendorf (2002, p. 141) states reliability is "paramount" in content analysis and content analytic results are "useless" without its establishment. As such, the above estimates of ICR's frequency are likely to exceed its prevalence in the broader qualitative literature.

The practice and evaluation of qualitative research is often guided by published checklists that stipulate steps that improve the quality of an analysis (Barbour, 2001). Some recommendation of multiple coding often appears on such checklists (Barbour, 2001). For instance, the Consolidated Criteria for Reporting Qualitative Studies requires specification of the number of coders (Tong et al., 2007) and the National Institute for Health and Care Excellence (2012) quality appraisal guidelines query whether the analysis was reliable and whether data were coded by multiple people. Additionally, some peer-reviewed journals (e.g., *Social Science & Medicine, Journal of the Society for Social Work and Research, Journal of Nutrition Education & Behavior*) publish criteria for authoring and reviewing qualitative studies that include recommendations of ICR (Wu et al., 2016). ICR may therefore assist with achieving certain dissemination and impact pathways.

The inclusion of ICR in such quality criteria may suggest that in certain scholarly communities, ICR has become mainstreamed as a standard and expected step in qualitative analysis. Feng's (2014) study suggests ICR became more commonly reported in the 2000s, although inappropriate statistical procedures also grew around this time. The past decade has seen a general movement from calculation of basic percentage agreement, which statisticians agree is an inadequate index (Cohen, 1960; Hallgren, 2012; Lombard et al., 2002), toward more formal statistical tests such as Krippendorff's α (Feng, 2014). However, there remains considerable dissensus regarding the most effective way to conduct ICR assessment and more fundamentally regarding its propriety within a qualitative paradigm. The following sections review the arguments commonly raised in favor of and against the inclusion of ICR assessment in qualitative analysis.

Arguments in Favor of ICR

The most commonly cited rationale for performing an ICR assessment is to assess the rigor and transparency of the coding frame and its application to the data (Hruschka et al., 2004; Joffe & Yardley, 2003; MacPhail et al., 2016; Mays & Pope, 1995). Achieving high ICR can satisfy the research team and audience that the coding frame is sufficiently well specified to allow for its communicability across persons (Joffe & Yardley, 2003). For example, in cross-cultural studies of lay responses to earthquakes (Joffe et al., 2013) and HIV/AIDS (Joffe, 1999), ICR assessment provided confidence that data collected in different languages and cultural contexts was consistently coded, allowing for exploration of similarities and differences across cultural data sets. Although qualitative research, by definition, places value in the analyst's interpretation of data, the ultimate purpose of doing and publishing research is to share it with others (Yardley, 2008). ICR helps qualitative research achieve this communicative function by showing the basic analytic structure has meaning that extends beyond an individual researcher. The logic is that if separate individuals converge on the same interpretation of the data, it implies "that the patterns in the latent content must be fairly robust and that if the readers themselves were to code the same content, they too would make the same judgments" (Potter & Levine-Donnerstein, 1999, p. 266). Performing an ICR assessment ensures multiple individuals can understand and contribute to the analytic process. ICR therefore provides confidence that the analysis transcends the imagination of a single individual (Kurasaki, 2000).

One undeniably important element of ICR is an external quality-signaling function. Reporting ICR can help persuade readers that the analysis was performed conscientiously and consistently (Kurasaki, 2000). ICR can thus serve as a badge of trustworthiness. Indeed, some journal editors and reviewers may request or require a measure of ICR before agreeing to publish qualitative studies (Wu et al., 2016). Given qualitative research is still viewed with suspicion in some quarters, a concrete quality indicator demonstrating the rigor of the research procedure can greatly assist researchers in increasing the reach and influence of their research. This may be particularly welcome when communicating research to multidisciplinary audiences who may not be familiar with qualitative analysis (Hruschka et al., 2004).

This said, ICR need not be undertaken for exclusively extrinsic concerns. In many researchers' experience, the primary advantages of ICR are internal to the research process (Barbour, 2001; MacPhail et al., 2016). First, ICR motivates researchers to ensure consistency in coding decisions. This is important when data coding is distributed across multiple researchers, as large projects frequently necessitate (Burla et al., 2008; MacPhail et al., 2016). It is especially critical for cross-cultural or cross-linguistic studies, as in the studies by Joffe (1999; Joffe et al., 2013) mentioned above. ICR ensures workloads can be shared without compromising the internal cohesion of the analysis. Additionally, even at an intracoder

level, awareness that one's coding will be compared to that of others can provide an incentive to maintain high coding standards. Coding can be monotonous, and it is easy for the coder's mind to drift. The self-disciplinary function of embedding some monitoring into the process should not be underestimated.

Second, ICR fosters reflexivity and dialogue within the research team. Echoing Barbour (2001), the content of intercoder disagreements can be equally, if not more valuable than the ultimate degree of consistency. Any inconsistencies the ICR process reveals should be discussed among coders to clarify the conflicting interpretations responsible. These discussions should inform the refinement of the coding frame to improve precision (Joffe & Yardley, 2003). For example, early ICR assessment in an interview study of laypeople's associations with neuroscience (O'Connor & Joffe, 2014b) identified numerous codes that, while reflecting basic concepts in the theoretical framework used by the primary researcher, were not clear to an external researcher recruited to second-code the data (e.g., "self-control," "causal attribution"). This revelation impelled the revision of the coding frame to more tightly define the focus and boundaries of these conceptual codes. The iterative developments prompted by ICR mean that the ICR process itself is often more valuable than the final scores (MacPhail et al., 2016). The benefits of discussion between researchers are acknowledged by many researchers who opt not to quantitatively measure ICR, yet include a phase of informal intercoder comparison and discussion. Such collaborative exercises are undoubtedly intrinsically beneficial, but formally computing ICR makes these discussions more systematic and informed, by revealing codes' relative reliability status and efficiently directing attention to the specific codes proving ambiguous. The result is a more explicit and well-defined coding frame (Joffe & Yardley, 2003), which is the primary tool for analyzing the data.

Finally, although many qualitative studies are purely exploratory, some have tangible, real-world repercussions. For instance, Hruschka et al. (2004) describe qualitative studies undertaken to inform policy in the Centers for Disease Control and Prevention, which must be acted on by policy-makers from multidisciplinary backgrounds. As another example, an interview study by O'Connor and McNicholas (2019) was used to derive recommendations for how clinicians should communicate psychiatric diagnoses to young people and their families. Qualitative findings may influence governmental or organizational policy, case-specific decisions in medical or legal contexts, or the distribution of public or charitable funds. With such consequences at stake, any effort to increase confidence in the evidence-base is welcome.

Objections to ICR

ICR is by no means universally accepted as beneficial for qualitative studies. Perhaps the most frequently aired objection is that ICR essentially contradicts the interpretative agenda of qualitative research (Braun & Clarke, 2013; Hollway & Jefferson, 2013; Vidich & Lyman, 1994; Yardley, 2000). Much of this relates to the epistemological status of reliability within the qualitative tradition. In quantitative research, reliability relates to the stability of findings across time, contexts, and research instruments. By the logic of positivist research, if a finding is reliably substantiated across these dimensions, it is more likely to represent an objectively "true" phenomenon rather than an artefact of the research process (Bauer et al., 2000). In contrast, most qualitative epistemologies reject the notion of a single, objective, external "reality" the scientific method can directly reveal. Instead, qualitative scholars see their research field as composed of multiple perspectival realities that are intrinsically constituted by an individual's social context and personal history (Bauer et al., 2000). Qualitative researchers' role is not to reveal universal objective facts but to apply their theoretical expertise to interpret and communicate the diversity of perspectives on a given topic. Within this epistemological framework, researcher reflexivity and active personal engagement with the data are resources, not "noise" to be minimized (Yardley, 2008).

Inarguably, complete objectivity is not a realistic expectation while coding latent content (Potter & Levine-Donnerstein, 1999) and indeed may not be a desirable one. Krippendorf (2004) notes textual meanings only arise in the process of somebody conceptually engaging with them; some degree of interpretation is therefore necessary to discern the meaning a particular segment of text holds. Affirming the analytic necessity of interpretation does not, however, negate the possibility of producing an analysis that is systematic, explicit, and transparent (Bauer, 2000). Qualitative researchers have proposed alternative quality criteria that substitute the typical standards used in evaluating quantitative research (Bauer et al., 2000; Popay et al., 1998; Seale & Silverman, 1997; Yardley, 2000). Along with ICR, these include transparent reporting of the analytic procedures, producing "thick description" with plentiful samples of raw data, triangulation between numerous studies, attention to deviant cases, and asking research participants to validate the legitimacy of analytic interpretations.

Some researchers strongly object to the inclusion of ICR in qualitative analysis because they see it as an unwarranted attempt to import standards derived for positivist research (Guba & Lincoln, 1994; Madill et al., 2000). For instance, Stenbacka (2001, p. 552) states, "reliability has no relevance in qualitative research, where it is impossible to differentiate between researcher and method." Likewise, Braun and Clarke (2013) assert reliability is not an appropriate criterion for judging qualitative work and that quantitative measures of ICR are epistemologically problematic.

It is possible that some of the antagonism toward ICR arises from the mere word "reliability" and its conventional association with a quantitative paradigm. Indeed, some critics of ICR suggest alternative concepts such as "dependability" or "trustworthiness" may be acceptable (Braun & Clarke, 2013). It can be argued that the aim of attaining acceptable ICR does not *necessarily* imply there is a single true meaning inherent in the data, which is the concern underpinning most

epistemological objections to ICR (Braun & Clarke, 2013). Rather, it shows that a group of researchers working within a common conceptual framework can reach a consensual interpretation of the data. While this can be trivialized as merely proving that different researchers can be trained to interpret data in similar ways (Joffe & Yardley, 2003; Yardley, 2000), this in itself is not an insignificant achievement. As Yardley (2008) acknowledges, a wholesale rejection of any transferability of qualitative findings is unproductive: if findings were entirely idiosyncratic to individual studies, there would be little point in doing qualitative research. Indeed, valorizing the sanctity of the analyst's unique interpretation could be read as highly individualistic—which is ironic, given much qualitative research orients to a social constructionist epistemology. Within an intellectual community, it should be possible to develop confidence that researchers are analyzing data using a common conceptual framework. ICR is one way of establishing, rather than just assuming colleagues are understanding and using conceptual tools in similar ways.

An arguably more significant risk of ICR is the false precision that numerical information can convey. Research shows inclusion of entirely nonsensical mathematical information can inflate judgments of the quality of research reports (Eriksson, 2012). Merely including a high ICR figure may lead to unjustifiably positive judgments of otherwise weak studies. The skewing effect of quantitative information is particularly problematic for students or newcomers to qualitative research. In the authors' experience of teaching thematic analysis in undergraduate psychology programs, students who initially struggle with the open-ended nature of qualitative analysis can become disproportionately fixated on achieving satisfactory ICR at the expense of the substantive analytic work. It might therefore be advisable to defer teaching ICR until students are more comfortable with the core tenets of qualitative analysis. However, for an experienced qualitative researcher, incorporating a numerical measure of ICR need not compromise analytic depth. Additionally, many qualitative analyses draw on quantitative information, such as frequency counts of the number of interviews that contain a given code (Maxwell, 2010). If researchers intend to include numerical information in the analytic results, validating that through initial ICR assessment is good practice.

While the costs and benefits of ICR are inevitably specific to the research context in question, in the authors' experience the gains usually outweigh the risks. However, it is important to maintain perspective in relation to one's research questions and prevent ICR from becoming the focal point of an analysis. ICR is never an end in itself; it is merely a means to the ultimate goal of achieving an insightful and robust qualitative analysis.

Practical Considerations: Performing an ICR Assessment

Manual or electronic?. With today's technological resources, the coding process is greatly aided by specialized qualitative analysis software packages, particularly when large quantities of data are involved. Some software packages, such as NVivo, Dedoose, and QDA Miner, contain an integrated ICR calculation tool. Others may not have an inbuilt ICR function, but coding patterns can be exported to external tools (e.g., the coding analysis toolkit at http://cat.ucsur.pitt.edu/). However, in the authors' experience, the process by which automated ICR tools calculate an ICR figure is often opaque and overly sensitive to inconsequential differences in coders' files (e.g., when coders have selected data units that differ by a mere punctuation mark). An alternative means of retaining control over the process is to export coding data from a qualitative platform to a statistical software package (e.g., SPSS) and calculate reliability there. The process the authors have devised to do this is described in detail below.

Despite the popularity of qualitative software packages, some researchers prefer or are forced by resource constraints to perform analyses manually. Hand-performed coding, aided by colored highlighters and sticky notes, does not preclude ICR calculation but will almost certainly make it more difficult. If a researcher wishes to perform ICR and has no access to specialist packages, performing the coding using a generic word processing package such as Microsoft Word (e.g., by indexing the codes through the "Comment" function) or tabulating assigned codes in a spreadsheet would make the intercoder comparison more efficient.

How many coders? A minimum of two independent coders is necessary to establish ICR. The inclusion of additional coders beyond this depends on the pragmatic resources and requirements of the specific project. For large data sets where coding must be divided between multiple researchers, it may be important to establish that all coders are applying the coding frame in consistent ways. In such cases, the addition of any new researcher may require a further ICR calculation to assess this individual's performance.

As Campbell et al. (2013) acknowledge, the reality of many qualitative research projects, particularly in early-career contexts, is that a single coder codes the majority of the data. In such cases, ICR can be obtained by recruiting an additional person to code a sample of the data. Once satisfactory reliability has been established, the primary researcher then proceeds to code the remaining data alone.

What proportion of the data should be multiply coded? While some studies apply multiple coding to the entire data set, resource constraints usually mean ICR is calculated on just a subset of the data. However, there is little consensus regarding the proportion of the data set that facilitates a trustworthy estimate of ICR (Campbell et al., 2013). Depending on the size of the data set, 10–25% of data units would be typical. It is important this subsample is selected randomly or using some other justifiable criteria (e.g., selecting a member of each group in a stratified sample) to ensure representativeness of the entire data set.

Rather than investing several days double-coding a sizable amount of data, only to reveal poor reliability caused by easily soluble issues with the coding frame, it may be judicious to first double-code a small amount of data (e.g., one interview).

Informal comparison of code patterns should reveal any obvious problems with code definitions or interpretations. The coding frame can then be refined before commencing the formal independent double-coding with the larger subset of data.

Some researchers opt to implement ICR testing across repeated rounds until satisfactory reliability is achieved (Campbell et al., 2013; Hruschka et al., 2004). For instance, in a study of a HIV prevention trial in South Africa, MacPhail et al. (2016) used a stepwise method that recalculated ICR and refined the coding frame after each individual transcript. Although resource intensive, this method helps improve reliability due to more opportunities to clarify code definitions and remove redundant codes. However, it is possible this reliability may simply reflect "interpretive convergence" (Hruschka et al., 2004) between this particular group of coders rather than any noticeable improvements in the transparency and external communicability of the coding frame.

What level of independence should the coders have? It is generally accepted that the physical double-coding should be performed independently without conferral between coders. However, advice differs regarding the level of interaction coders should have prior to commencing the coding. Some researchers, whose analytic approaches prioritize increasing the coding frame's external objectivity, recommend coders should be people external to the research team who had no role in designing the coding frame (Kolbe & Burnett, 1991). Such an approach requires consideration of ethical and data protection implications involved in passing raw data to an external individual.

However, as previously discussed, many researchers value ICR not as a measure of "objectivity" but as a means of reflexively improving the analysis by provoking dialogue between researchers. If one aim of performing an ICR check is to identify areas needing clarification, some discussion between coders is necessary to identify how and why interpretations conflict. In such cases, a first round of independent coding could be followed by a meeting where differences are discussed, the coding frame is revised, and a second round of independent coding commences (Campbell et al., 2013; Hruschka et al., 2004).

Whether prior training of coders is required depends on the depth of the analysis. If codes are simply categorizing surface-level features of data (e.g., whether a newspaper article is an opinion piece or news report), merely possessing a clearly specified coding frame should be sufficient to allow an entirely independent coder to commence. However, if the analysis involves coding more "latent" features of the data, which require greater degrees of interpretation, the coder may need training in relevant theoretical concepts. The issue of code depth is discussed in more detail below.

How should data be segmented? The specific data "units" or segments coded differ across studies depending on the research aims. At the broadest level, each source of data could be coded as a single data unit—for example, holistically coding a whole interview or entire media article as one entity. Other study protocols might instruct coders to segment data into smaller prespecified coding units, for example, each paragraph or each response to an interviewer's question. More fine-grained analyses may code each individual line or sentence. Finally, some studies will not prespecify any consistent data unit and instead code ad hoc segments the researcher determines to be conceptually meaningful; for example, a block of sentences that organically elaborate one cohesive idea. Figure 1 shows an example of the latter form of coding in ATLAS.ti, taken from the analysis reported in Joffe et al. (2013).

Each of these strategies has distinct strengths and weaknesses, which must be evaluated according to their coherence with the research aims. In general, larger units of analysis are associated with greater validity: the more data units' original contextualization is preserved, the more valid their interpretation. The meaning of a particular structural element (e.g., a sentence) can often be difficult to ascertain in isolation from its neighboring text. However, coding larger units invites an increased degree of complexity, as it is more likely they contain a range of different (sometimes contradictory) ideas. This poses a challenge when operating an "exclusive" coding strategy that allows for only one code to be assigned to each data unit, though it is less problematic when the protocol allows for coding with multiple codes. However, coding large data units often involves a compromise of analytic sensitivity: the linguistic nuances that are central to many qualitative questions can be lost. In relation to ICR specifically, longer text units are usually associated with poorer reliability (Hruschka et al., 2004).

The "unitization problem" is often a distinct challenge for researchers analyzing interview data (Campbell et al., 2013; Hollway & Jefferson, 2013). Interview data can often be rather unsystematic, with respondents taking variable time to communicate an idea and abruptly jumping between topics. To faithfully capture the meanings conveyed, an *ad hoc* data unitization strategy, where the researcher determines how to segment the transcript into conceptually meaningful quotes, is often appropriate. This has been the authors' preference in analyzing interview data in the past. However, this can cause challenges for ICR, as there is no guarantee that different coders will select the same quotes when applying codes. Without any predefined guidance regarding data units, some individual coders ("lumpers") orient toward selecting larger, more contextualized segments, while others ("splitters") apply codes more specifically to short segments. This can cause major difficulty in any form of automated ICR calculation: When coders' selection of text segments varies by a mere digit or punctuation mark, automated systems can record these as different data units and the calculated ICR coefficient is compromised (MacPhail et al., 2016).

Kurasaki (2000) proposes one strategy for managing this issue: allowing coders to select their own segments, then randomly picking certain lines in the document, and comparing codes recorded within a radius of five lines. In the authors' previous research, they have developed an alternative strategy very similar to one described by Campbell et al. (2013). This approach takes as its premise that consistency in where coders

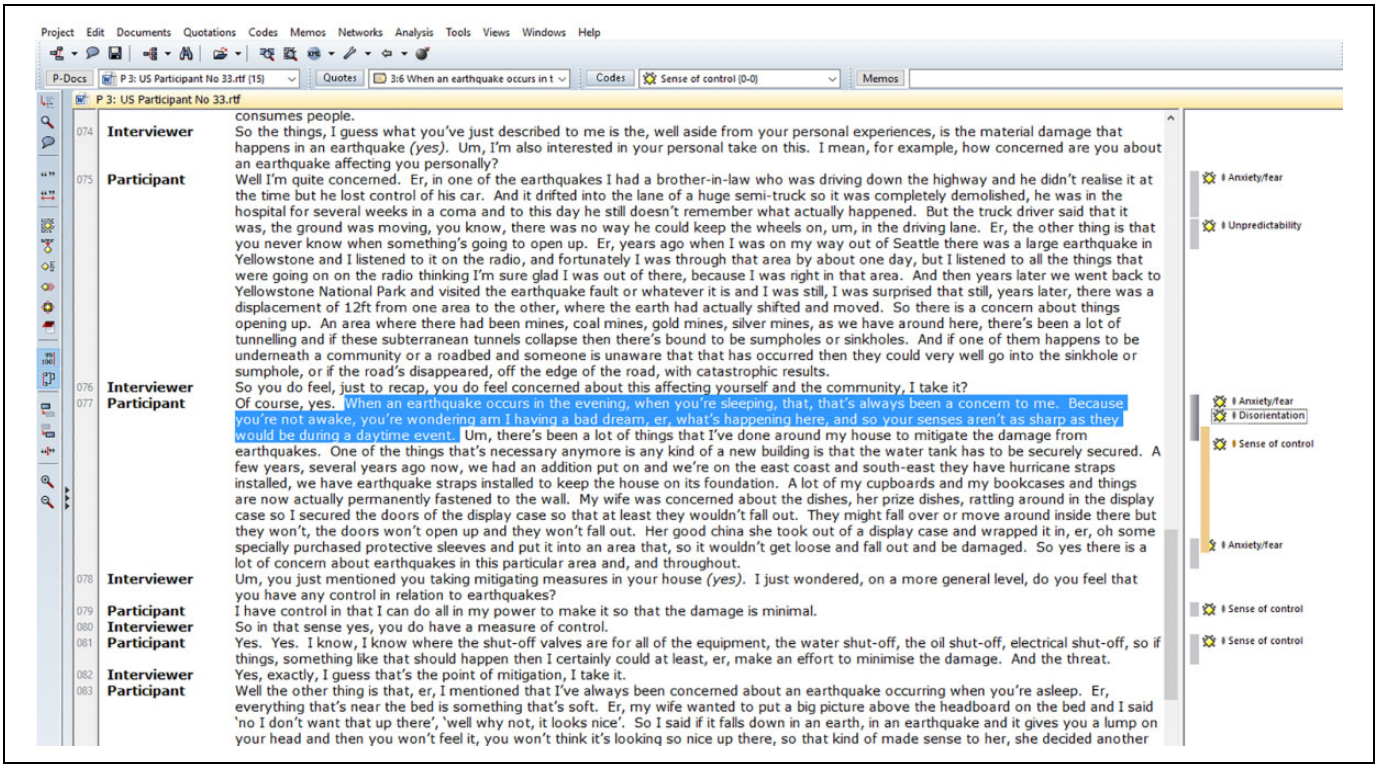

Figure 1. Example of coded interview data in ATLAS.ti (Joffe et al., 2013).

choose to start and end quotes is of minimal analytical significance[2]: more important is ensuring that when given a certain segment of text, similar codes are applied. In this method, one researcher first approaches the transcript, segmenting it as they see fit and applying relevant codes. Campbell et al. (2013) suggest this should be the principal investigator or person most familiar with the subject area, who is therefore more qualified to determine the "meaningful conceptual breaks" (Campbell et al., 2013, p. 304). Once the first coder has saved the coded transcript, they can then create a second document where the data segments are visible but the codes removed. Some qualitative software programs, such as ATLAS.ti, allow users to "unlink" the codes; this creates an uncoded file that can then easily be passed to a second coder.[3] Another option may be to highlight the relevant data segments in a text document. The second coder then uses their own judgment to code the segments they have received.

This example of an ICR-amenable coding strategy illustrates how the conditions necessary for ICR assessment can constrain coding practices. It is incumbent on individual researchers to determine whether the benefits ICR offers for a particular project outweigh the sacrifices of analytic flexibility involved.

How many codes? The number of codes in the coding frame must be dictated by the research questions and diversity of content within the data. A further variable impinging on code quantity relates to whether the analysis permits exclusive or multiple coding (i.e., whether the protocol stipulates that each data unit can have just one or multiple codes). Multiple coding is often necessary to authentically capture meaning in complex data such as interviews (Campbell et al., 2013) and usually inflates the total number of codes. In relation to ICR, it is worth being aware that the more codes are available, the lower ICR is likely to be (Hruschka et al., 2004; Roberts et al., 2019). This is because it is difficult for coders to familiarize themselves with a lengthy coding frame and hold all potential codes in their working memory when considering many data units. Additionally, very elaborate coding frames often include some codes with low frequency of occurrence, which may not meet the minimum number of observations required for certain reliability statistics to be performed. MacQueen et al. (1998) suggest researchers concerned with achieving satisfactory reliability should work with an upper limit of 30–40 codes. Hruschka et al. (2004) recommend a limit of approximately 20 and further suggest that for semistructured interview data, codes should be specific to particular interview questions. Another potential rule of thumb is to disallow more codes than there are interviews or other relevant data units.

These suggested upper limits should not be taken as dogma. As always, the analytic aims particular to each study should be the primary consideration in designing the analysis. Code comprehensiveness should not be sacrificed purely for the sake of achieving ICR. With enough time and attention, it is possible to implement ICR procedures that reduce the load on coders' cognitive resources. A well-structured, conceptually and visually clear coding frame minimizes the burden on coders' working memory. Campbell et al. (2013) additionally suggest

	Interview_ ID	DataUnit_ ID	Coder1__ Sadness	Coder2__ Sadness	Coder1__ Joy	Coder2__ Joy	Coder1__ Fear	Coder2__ Fear
1	1	1	0	0	1	0	0	0
2	1	2	0	0	1	1	0	0
3	1	3	0	0	0	0	0	0
4	1	4	0	0	0	0	1	1
5	1	5	1	1	0	0	0	0
6	1	6	0	0	0	0	0	1

Figure 2. Example of how both coders' decisions are represented in SPSS. The column variables include IDs for the interview and data unit (quote) in question, and both coders' decisions regarding applications of three emotion-related codes. Coding patterns are largely similar, except only Coder 1 applied joy to Data Unit 1, while only Coder 2 applied *fear* to Data Unit 6.

grouping codes into "families" of related codes and taking multiple family-specific "passes" at the data, coding according to one family of codes at a time. With complex coding frames, this absolves coders of the requirement to bear all codes in mind simultaneously.

How interpretative should codes be? Different research questions require different "depths" of coding. This relates to the study's level of interest in cataloguing "manifest" surface-level content or deeper "latent" meanings. Some coding will record unambiguous, purely factual data (e.g., geographical location). If coders are conscientious, such coding should be near-perfectly synchronous. Other codes will index descriptive information that requires some interpretation but should nevertheless be relatively apparent. For instance, O'Connor (2017) found high ICR when coders were asked to judge whether newspaper articles expressed a supportive, antagonistic, or neutral attitude toward same-sex marriage. In developing even straightforward codes, researchers should avoid assuming anything is "obvious" (MacQueen et al., 1998): the more explicitly defined the codes, the more transparent the process and the higher ICR is likely to be (Joffe & Yardley, 2003).

Some qualitative studies involve coders deploying high levels of interpretation in coding latent features of the data, which may require familiarity with relevant theoretical concepts. More conceptually sophisticated coding frames typically produce lower ICR calculations. However, this consideration alone should not deter researchers from including more interpretative codes in their coding frame: consistency with study aims should always be the primary consideration. Theoretical relevance or meaning should never be sacrificed for reliability (Hruschka et al., 2004). To minimize confusion, a coding frame can include not only examples of typical manifestations of a complex code but also specify its qualifications and exclusions (e.g., "this code does *not* apply to instances where . . . "; Boyatzis, 1998; Roberts et al., 2019).

How should ICR be calculated? Numerous measures of ICR are available. Previous reviews of the literature indicate the most common method is simply reporting the percentage of data units on which coders agree (Feng, 2014; Kolbe & Burnett, 1991). Miles and Huberman (1994) suggest reliability can be calculated by dividing the number of agreements by the total number of agreements plus disagreements. However, percentage-based approaches are almost universally rejected as inappropriate by methodologists because percentage figures are inflated by some agreement occurring by chance (Cohen, 1960; Hallgren, 2012; Lombard et al., 2002). Additionally, while the percentage agreement approach appeals to researchers due to its apparently straightforward manual calculation, attempts to perform this calculation can reveal unanticipated complexities. This occurs especially when the protocol allows for multiple coding of data units. If one coder has applied three codes to a piece of text, and another coder has applied four, with two codes overlapping between coders, it is not obvious how that should be quantified. The procedure becomes even more complex if there are more than two coders (McHugh, 2012).

Statistical tests developed for measuring ICR include Cohen's kappa, Krippendorff's alpha, Scott's pi, Fleiss' K, Analysis of Variance binary ICC, and the Kuder-Richardson 20. The statistical foundations of these measures are beyond the scope of this article but are fully discussed elsewhere (Banerjee et al., 1999; Davey et al., 2010; Feng, 2013; Hallgren, 2012; Hayes & Krippendorff, 2007; Rust & Cooil, 1994). The primary advantage these statistics offer over percentage agreement is correction for the probability a certain amount of agreement occurs by chance. These tests can also be used to assess the reliability of codes that have been applied nonexclusively (i.e., multiple codes applied to a single data segment). Krippendorff's alpha appears to be increasing in popularity (Feng, 2014) and is often preferred for its flexibility: it can incorporate more than two coders and incorporate ordinal, interval and ratio as well as nominal data (Lombard et al., 2002). Chi square, Cronbach's alpha and correlational tests such as Pearson's r are not appropriate measures of ICR (Lombard et al., 2002).

While it is possible to perform such analyses by hand, most contemporary researchers rely on algorithms embedded in their qualitative analysis or statistical software. The authors of this article have generally adopted the strategy of exporting each coder's coded data from ATLAS.ti into SPSS. The SPSS files generated present each data unit as a row and each code as a

column (see example in Figure 2). If a code has been applied to a data unit, the relevant cell shows 1, and if that code has not been applied, the cell records 0. If the two coders' ATLAS.ti files were consistently structured in terms of codes available and text segments coded, the two SPSS files can be merged. In this merged data set, each code should have two corresponding columns representing both coders' applications of that code (the variable naming system should clearly indicate which coder is responsible for each column). SPSS' statistical functionalities can then be used to implement whatever statistical test has been chosen.

How should results be presented? It is relatively common to present ICR using a single pooled or average value that represents the reliability of the coding frame as a unitary instrument (Feng, 2014). For example, Burla et al.'s (2008) analysis of experiences of low back pain reports a single kappa statistic that encompasses the coding frame as a whole. This gives a concise way of summarizing the results of the ICR process.

However, caution should be exercised regarding decisions to collapse ICR into a single summary statistic. First, if the aim of ICR is to improve the coding frame, assessing reliability on a code-specific level is critical to identify codes that require refinement. Second, pooling all codes' reliability figures means codes with poor reliability can be "hidden" or canceled out by codes that perform very well (usually because they are codifying more straightforward manifest content). Nevertheless, presenting the reliability figure for each individual code remains infrequent in published reports (Feng, 2014), perhaps due to space constraints. A parsimonious alternative may be to present the range and distribution of all codes' ICR performance, perhaps with supplemental individual coefficients in an appendix.

How should results be interpreted? There is considerable inconsistency in the interpretation of ICR results. Mere statistical significance is never an acceptable indication of ICR: understanding results requires interpretation of the coefficient in question. For percentage agreement approaches, there is no universally accepted threshold for what indicates acceptable reliability, but Miles and Huberman (1994) suggest a standard of 80% agreement on 95% of codes. Most of the commonly used statistical tests of ICR present results on a scale between -1 to $+1$, with figures closer to 1 indicating greater correspondence. Neuendorf (2002) reviews "rules of thumb" that exist for interpreting ICR values, observing ICR figures over .9 are acceptable by all, and over .8 acceptable by many, but considerable disagreement below that. Researchers often cite Landis and Koch's (1977) recommendation of interpreting values less than 0 as indicating no, between 0 and 0.20 as slight, 0.21 and 0.40 as fair, 0.41 and 0.60 as moderate, 0.61 and 0.80 as substantial, and 0.81 and 1 as nearly perfect agreement.

All such guidelines are ultimately arbitrary, and the researcher must judge what represents acceptable agreement for a particular study. Studies that influence important medical, policy, or financial decisions arguably merit a higher ICR threshold than exploratory academic research (Hruschka et al., 2004; Lombard et al., 2002). For instance, McHugh (2012) proposes a more conservative system of acceptability thresholds when using Cohen's kappa coefficients in the context of clinical decision-making. Whatever interpretative framework is chosen should be stipulated in advance and not decided *post hoc* after results are viewed.

How should results be acted on? Again, there is no universal agreement regarding how to manage low-performing codes. Much depends on whether the researchers are approaching ICR as a one-off validation of the coding process or as a tool through which the coding frame can be progressively improved. Some researchers may opt to discard codes below a certain ICR threshold. Others may modify poorly performing codes and double-code a further sample of data with the revised coding frame, repeating this process until an acceptable ICR is attained. Others may judge that ICR's utility is purely in refining a theoretically dictated coding frame and that ICR results should not inform decisions about retaining or removing codes selected for their conceptual importance. The appropriateness of any such approach can only be judged in relation to the specific research aims and context.

Researchers must also decide how to treat instances of intercoder disagreement when finalizing the "definitive" coded data set. Some research teams may introduce a third coder and adopt a "majority rules" decision. Others may decide the judgments of one coder (usually the PI or more experienced researcher) outweigh those of the other. Finally, some research teams may adopt a consensus approach where disagreements are discussed and joint decisions reached. Campbell et al. (2013) describe such a strategy of "negotiated agreement" (Campbell et al., 2013, p. 305), which ultimately increased reliability from 54% to 96%.

Once the coding frame is finalized, it should be systematically applied to the entire data set. This typically involves the recoding of data originally coded during the ICR process (MacQueen et al., 1998).

Suggested Procedure for ICR Assessment

The preceding section lays out the decisions that must be taken when designing an ICR assessment within qualitative research. The advantages and drawbacks of the various options will necessarily be relative to the specific research question and context, and the researcher must decide and justify which options are most appropriate.

Figure 3 presents a suggested procedure for the various steps of ICR assessment. Before beginning the coding, the researchers must make *a priori* decisions regarding the number of coders, amount of data that will be coded in duplicate, the unit of coding (i.e., sentences, paragraphs, conceptually meaningful "chunks"), the conceptual depth that codes will capture, the reliability measure that will be calculated, and the threshold that will indicate acceptable reliability. These decisions are necessarily project-specific and should be dictated by the research aims rather than the proximate goal of attaining acceptable reliability.

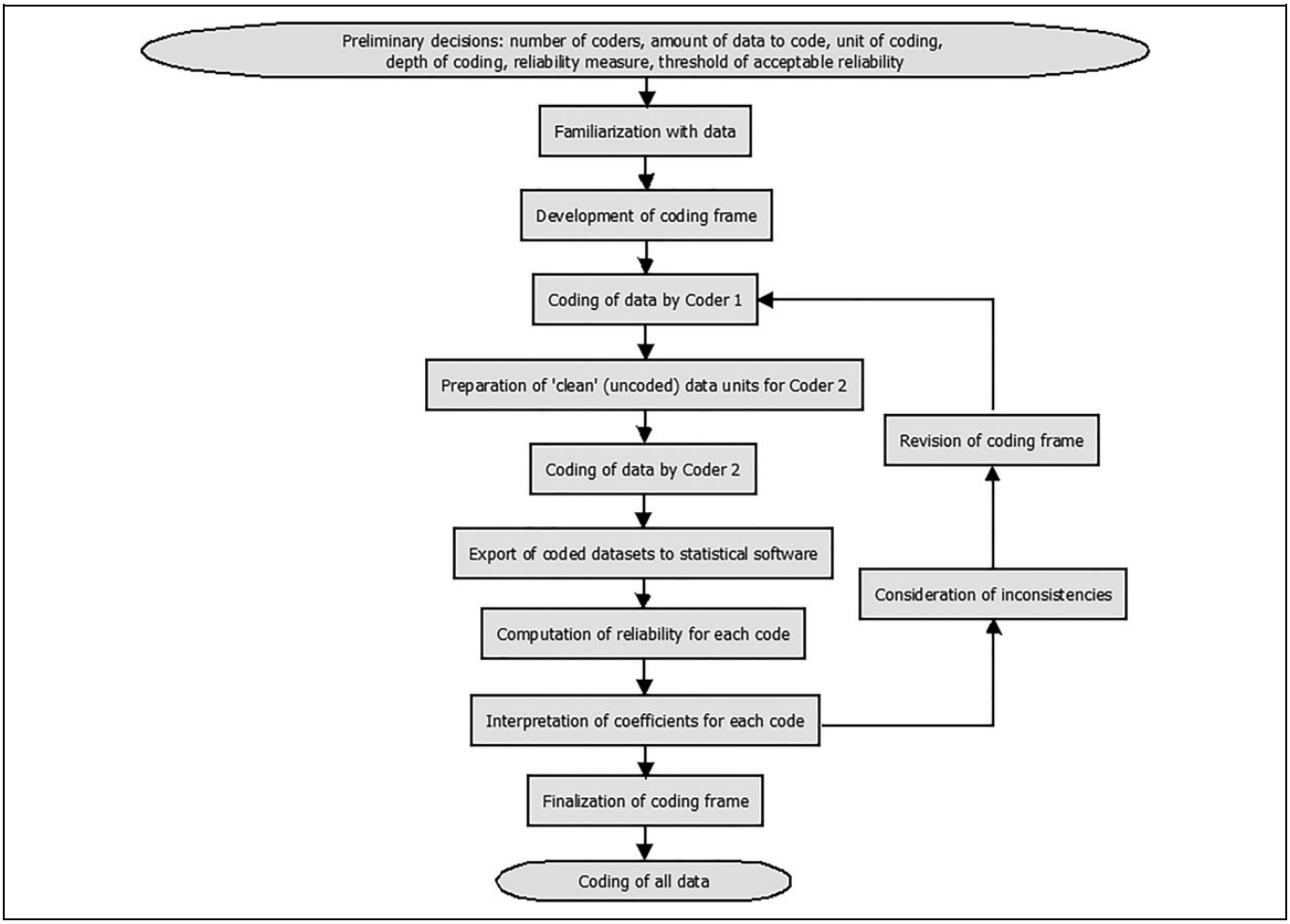

Figure 3. Suggested procedure for intercoder reliability assessment.

The practical process of coding begins with immersion in the data, usually through intensive reading. Through this familiarization, the research team develops a first draft of a coding frame that may, depending on the project, contain either or both inductive and deductive codes. In this suggested procedure, the first coder then applies this coding frame to the data, ideally using a qualitative software package. Coder 1 moves through the subset of data included in the ICR test, segmenting the data into data units and labeling them with relevant codes. Once complete, the coded file is saved. Coder 1 then duplicates the file, removes the code names they have assigned, and passes to Coder 2 a "clean" file that displays the breaks indicating the data units, but not their associated codes. Coder 2 then uses the coding frame to independently code the data units that are visible on the cleaned file. It may be beneficial to first informally compare coding on a small quantity of data (e.g., one interview), to clarify any immediately apparent code misinterpretations before formal reliability evaluation begins.

To compute reliability, both coded files can be exported into a statistical program such as SPSS (some qualitative packages allow files to be converted automatically; others require initial conversion to a CSV file). The two statistical data files can then be merged so they appear as illustrated in Figure 2. The functionalities of the statistical software can be used to compute the reliability statistic of interest for each code in the coding frame. Results should be interpreted according to the *a priori* threshold of acceptable reliability. Codes that fall short of the threshold can be evaluated to identify potential reasons for inconsistency of interpretation, and removed or revised in accordance with the team's best judgment. The revised coding frame can be evaluated using the same process, preferably on a different subset of data. Once the research team is satisfied with the overall reliability of the coding frame, the entire data set can be coded by a single coder or team of coders.

Conclusion

The value of qualitative research lies in its sensitivity to the diverse meanings people derive of particular issues within particular contexts: method and analysis can and should be adapted to suit the specific features of the phenomenon under investigation. This makes it difficult to generate one-size-fits-all guidelines. However, a consensus has developed regarding the value of maintaining a set of quality criteria that help

researchers design their studies and help audiences differentiate high- from low-quality research (Bauer et al., 2000; O'Brien et al., 2014; Popay et al., 1998; Seale & Silverman, 1997; Yardley, 2000). ICR is one candidate quality criterion. It may not be appropriate for every qualitative study and is not a "magic bullet" for those studies that do include it. ICR attests to the robustness of the coding process, which structures the entire subsequent analysis. However, it is no guarantee of the trustworthiness of either prior data collection and preparation or subsequent theme generation and reporting. A broad sensitivity to accepted principles and practices in the qualitative tradition remains paramount.

In appropriate research contexts, ICR assessment can improve both the internal quality and external reception of qualitative studies. Key benefits include improving the systematicity, communicability, and transparency of the coding process; promoting reflexivity and dialogue within research teams; and helping to satisfy diverse audiences of the trustworthiness of the research. By collating the key arguments for and against ICR and outlining the practical requirements of performing it, this article endeavors to equip researchers make informed decisions about whether and how to incorporate ICR assessment into their analyses.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Dr. O'Connor's work on this article was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 702970.

ORCID iD

Cliodhna O'Connor https://orcid.org/0000-0001-8134-075X

Notes

1. As the aim of content analysis is typically to move "from words to numbers" (Franzosi, 2004) by producing frequency counts of features of textual data, its characterization as a form of qualitative analysis can be questioned. However, even a content analysis whose output is entirely numerical is punctured by qualitative processes at several points: Reading is a fundamentally qualitative activity (Krippendorf, 2004), as is the discerning of the qualities and distinctions of the categories to be counted (Bauer, 2000), and the assigning of codes to particular data segments. Most authorities on content analysis therefore characterize it as an approach that bridges the qualitative–quantitative divide (Bauer, 2000; Hsieh & Shannon, 2005; Krippendorf, 2004).
2. It is acknowledged that some researchers will disagree with this; this is simply a position consistent with the authors' own theoretical and methodological commitments. This strategy will not be suitable for all qualitative studies.
3. To envision what this file looks like in relation to Figure 1, imagine the code names in the right-hand margin have disappeared. By clicking on the shaded bars on the right, the second coder can see the segments of text that have been coded, but not the particular codes that the first coder applied.

References

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: An empirical study. *Sociology*, *31*, 597–606. https://doi.org/10.1177/0038038597031003015

Attride-Stirling, J. (2001). Thematic networks: An analytic tool for qualitative research. *Qualitative Research*, *1*, 385–405. https://doi.org/10.1177/146879410100100307

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*, 3–23. https://doi.org/10.2307/3315487

Barbour, R. S. (2001). Checklists for improving rigour in qualitative research: A case of the tail wagging the dog? *British Medical Journal*, *322*, 1115–1117. https://doi.org/10.1136/bmj.322.7294.1115

Bauer, M. W. (2000). Classical content analysis: A review. In M. W. Bauer & G. Gaskell (Eds.), *Qualitative researching with text, image and sound: A practical handbook* (pp. 131–151). Sage.

Bauer, M. W., Gaskell, G., & Allum, N. C. (2000). Quality, quantity and knowledge interests: Avoiding confusions. In M. W. Bauer & G. Gaskell (Eds.), *Qualitative researching with text, image and sound: A practical handbook* (pp. 3–17). Sage.

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2013). *Successful qualitative research*. Sage.

Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., & Abel, T. (2008). From text to codings: Intercoder reliability assessment in qualitative content analysis. *Nursing Research*, *57*, 113. https://doi.org/10.1097/01.NNR.0000313482.33917.7d

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, *42*, 294–320. https://doi.org/10.1177/0049124113500475

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

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

Davey, J. W., Gugiu, P. C., & Coryn, C. L. S. (2010). Quantitative methods for estimating the reliability of qualitative data. *Journal of MultiDisciplinary Evaluation*, *6*, 140–162.

Eriksson, K. (2012). The nonsense math effect. *Judgment and Decision Making*, *7*, 746–749.

Feng, G. C. (2013). Factors affecting intercoder reliability: A Monte Carlo experiment. *Quality & Quantity*, *47*, 2959–2982. https://doi.org/10.1007/s11135-012-9745-9

Feng, G. C. (2014). Intercoder reliability indices: Disuse, misuse, and abuse. *Quality & Quantity*, *48*, 1803–1815. https://doi.org/10.1007/s11135-013-9956-8

Franzosi, R. (2004). *From words to numbers: Narrative, data and social science*. Cambridge University Press.

Gaskell, G. (2000). Individual and group interviewing. In M. W. Bauer & G. Gaskell (Eds.), *Qualitative Researching with Text, Image and Sound: A Practical Handbook* (pp. 38–56). Sage.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105–117). Sage.

Hallgren, K. A. (2012). Computing inter-rater reliability for observaional data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23–34.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77–89. https://doi.org/10.1080/19312450709336664

Hollway, W., & Jefferson, T. (2013). *Doing qualitative research differently: A psychosocial approach* (2nd ed.). Sage.

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*, 307–331. https://doi.org/10.1177/1525822X04266540

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*, 1277–1288. https://doi.org/10.1177/1049732305276687

Joffe, H. (1999). *Risk and 'the other'*. Cambridge University Press.

Joffe, H. (2012). Thematic analysis. In D. Harper & A. Thompson (Eds.), *Qualitative research methods in mental health and psychotherapy: An introduction for students and practitioners* (pp. 209–223). Wiley-Blackwell.

Joffe, H., & Elsey, J. (2014). Free association in Psychology and the Grid Elaboration Method. *Review of General Psychology*, *18*, 173–185. https://doi.org/10.1037/gpr0000014

Joffe, H., Rossetto, T., Bradley, C., & O'Connor, C. (2018). Stigma in science: The case of earthquake prediction. *Disasters*, *42*, 81–100. https://doi.org/10.1111/disa.12237

Joffe, H., Rossetto, T., Solberg, C., & O'Connor, C. (2013). Social representations of earthquakes: A study of people living in three highly seismic areas. *Earthquake Spectra*, *29*, 367–397. https://doi.org/10.1193/1.4000138

Joffe, H., & Smith, N. (2016). City dweller aspirations for cities of the future: how do environmental and personal wellbeing feature? *Cities*, *59*, 102–112. https://doi.org/10.1016/j.cities.2016.06.006

Joffe, H., Washer, P., & Solberg, C. (2011). Public engagement with emerging infectious disease: The case of MRSA in Britain. *Psychology & Health*, *26*, 667–683. https://doi.org/10.1080/08870441003763238

Joffe, H., & Yardley, L. (2003). Content and thematic analysis. In D. F. Marks & L. Yardley (Eds.), *Research methods for clinical and health psychology* (pp. 56–68). Sage.

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research*, *18*, 243–250.

Krippendorf, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage.

Kurasaki, K. S. (2000). Intercoder reliability for validating conclusions drawn from open-ended interview data. *Field Methods*, *12*, 179–194. https://doi.org/10.1177/1525822X0001200301

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Lavie-Ajayi, M., & Joffe, H. (2009). Social representations of female orgasm. *Journal of Health Psychology*, *14*, 98–107. https://doi.org/10.1177/1359105308097950

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*, 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

MacPhail, C., Khoza, N., Abler, L., & Ranganathan, M. (2016). Process guidelines for establishing intercoder reliability in qualitative studies. *Qualitative Research*, *16*, 198–212. https://doi.org/10.1177/1468794115577012

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

Madill, A., Jordan, A., & Shirley, C. (2000). Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies. *British Journal of Psychology*, *91*, 1–20. https://doi.org/10.1348/000712600161646

Maxwell, J. A. (2010). Using numbers in qualitative research. *Qualitative Inquiry*, *16*, 475–482. https://doi.org/10.1177/1077800410364740

Mays, N., & Pope, C. (1995). Rigour and qualitative research. *BMJ: British Medical Journal*, *311*, 109–112. https://doi.org/10.1136/bmj.311.6997.109

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

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.

National Institute for Health and Care Excellence. (2012). *Methods for the development of NICE public health guidance (third edition) | Guidance and guidelines | NICE*. https://www.nice.org.uk/process/pmg4/chapter/appendix-h-quality-appraisal-checklist-qualitative-studies

Neuendorf, K. A. (2002). *The content analysis guidebook*. Sage.

O'Brien, B. C., Harris, I. B., Beckman, T. J., Reed, D. A., & Cook, D. A. (2014). Standards for reporting qualitative research: A synthesis of recommendations. *Academic Medicine: Journal of the Association of American Medical Colleges*, *89*, 1245–1251. https://doi.org/10.1097/ACM.0000000000000388

O'Connor, C. (2012). Using social representations theory to examine lay explanation of contemporary social crises: The case of Ireland's recession. *Journal of Community & Applied Social Psychology*, *22*, 453–469. https://doi.org/10.1002/casp.1125

O'Connor, C. (2017). 'Appeals to nature' in marriage equality debates: A content analysis of newspaper and social media

discourse. *British Journal of Social Psychology*, *56*, 493–514. https://doi.org/10.1111/bjso.12191

O'Connor, C., & Joffe, H. (2013). Media representations of early human development: Protecting, feeding and loving the developing brain. *Social Science & Medicine*, *97*, 297–306. https://doi.org/10.1016/j.socscimed.2012.09.048

O'Connor, C., & Joffe, H. (2014a). Gender on the brain: A case study of science communication in the new media environment. *PLoS One*, *9*, e110830. https://doi.org/10.1371/journal.pone.0110830

O'Connor, C., & Joffe, H. (2014b). Social representations of brain research exploring public (dis)engagement with contemporary neuroscience. *Science Communication*, *36*, 617–645. https://doi.org/10.1177/1075547014549481

O'Connor, C., & Joffe, H. (2015). How the public engages with brain optimization the media-mind relationship. *Science, Technology & Human Values*, *40*, 712–743. https://doi.org/10.1177/0162243915576374

O'Connor, C., Kadianaki, I., Maunder, K., & McNicholas, F. (2018). How does psychiatric diagnosis affect young people's sense of self and social identity? A systematic review and synthesis of the qualitative literature. *Social Science & Medicine*, *212*, 94–119. https://dx.doi.org/10.1016/j.socscimed.2018.07.011

O'Connor, C., & McNicholas, F. (2019). *"Plopped into a different universe": The lived experience of diagnostic shifts in child and adolescent mental health contexts*. Manuscript submitted for publication.

O'Connor, C., Rees, G., & Joffe, H. (2012). Neuroscience in the public sphere. *Neuron*, *74*, 220–226. https://doi.org/10.1016/j.neuron.2012.04.004

Oktay, J. S. (2012). *Grounded theory*. Oxford University Press.

Popay, J., Rogers, A., & Williams, G. (1998). Rationale and standards for the systematic review of qualitative literature in health services research. *Qualitative Health Research*, *8*, 341–351. https://doi.org/10.1177/104973239800800305

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*, 258–284. https://doi.org/10.1080/00909889909365539

Roberts, K., Dowell, A., & Nie, J.-B. (2019). Attempting rigour and replicability in thematic analysis of qualitative research data; A case study of codebook development. *BMC Medical Research Methodology*, *19*, 66. https://doi.org/10.1186/s12874-019-0707-y

Rust, R. T., & Cooil, B. (1994). Reliability measures for qualitative data: Theory and implications. *Journal of Marketing Research*, *31*, 1–14. https://doi.org/10.2307/3151942

Seale, C., & Silverman, D. (1997). Ensuring rigour in qualitative research. *European Journal of Public Health*, *7*, 379–384. https://doi.org/10.1093/eurpub/7.4.379

Smith, J. A., Jarman, M., & Osborn, M. (1999). Doing interpretative phenomenological analysis. In M. Murray & K. Chamberlaine (Eds.), *Qualitative health psychology: Theories and methods* (pp. 218–240). Sage. https://doi.org/10.4135/9781446217870

Smith, N., & Joffe, H. (2009). Climate change in the British press: The role of the visual. *Journal of Risk Research*, *12*, 647–663. https://doi.org/10.1080/13669870802586512

Smith, N., & Joffe, H. (2013). How the public engages with global warming: A social representations approach. *Public Understanding of Science*, *22*, 16–32. https://doi.org/10.1177/0963662512440913

Stenbacka, C. (2001). Qualitative research requires quality concepts of its own. *Management Decision*, *39*, 551–556. https://doi.org/10.1108/EUM0000000005801

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*, 45. https://doi.org/10.1186/1471-2288-8-45

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, *19*, 349–357. https://doi.org/10.1093/intqhc/mzm042

Vidich, A. J., & Lyman, S. M. (1994). Qualitative methods: Their history in sociology and anthropology. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 23–59). Sage.

Wu, S., Wyant, D. C., & Fraser, M. W. (2016). Author guidelines for manuscripts reporting on qualitative research. *Journal of the Society for Social Work and Research*, *7*, 405–425. https://doi.org/10.1086/685816

Yardley, L. (2000). Dilemmas in qualitative health research. *Psychology & Health*, *15*, 215–228. https://doi.org/10.1080/08870440008400302

Yardley, L. (2008). Demonstrating validity in qualitative psychology. In J. A. Smith (Ed.), *Qualitative psychology: A practical guide to research methods* (pp. 235–251). https://eprints.soton.ac.uk/54781/