American Journal of Qualitative Research
2024, Vol. 8 No. 3, pp. 23-43
https://doi.org/10.29333/ajqr/14487
© 2024 AJQR. http://www.ajqr.org

AMERICAN JOURNAL OF QUALITATIVE RESEARCH

ISSN: 2576-2141

Inter-Coder Agreement in Qualitative Coding: Considerations for its Use

Sean N. Halpin[1]
RTI-International, North Carolina, USA

ABSTRACT
The historically quantitative-dominated field of health sciences has increasingly embraced qualitative methods. However, calls for quantitative measures of rigor, such as Inter-coder Agreement (ICA), remain. The aim of this manuscript is to demystify ICA and provide practical guidance. I begin by describing considerations while planning for ICA, including differences between various ICA tests (i.e., percent agreement, Holsti Method, Cohen's kappa, Krippendorf's alpha, and Gwet's AC1 and AC2), setting the threshold of acceptability for your chosen test, deciding whether to use qualitative data analysis software, choosing the number of coders, selecting what data will be coded by more than one coder, developing a deductive codebook, creating a process for resolving coding disagreements, and establishing an audit trail for codebook changes. Next, I provide step-by-step guidance on an iterative process used for enacting ICA. Finally, I discuss the importance of reporting, emphasizing clarity, conciseness, completeness, and accuracy.

KEYWORDS: Trustworthiness, rigor, inter-rater reliability, qualitative coding

Historically, quantitative methods have dominated the health sciences. However, with the increasing recognition of the need for patient-centered research approaches in the health sciences, qualitative methods have proliferated (Creswell, 2003). Indeed, the United States Food and Drug Administration requires patient experience data using quantitative and qualitative methods to be included in all drug applications (Gabay, 2017). As such, traditional quantitative health scientists expect convincing proof of trustworthiness, including a quantitative measure of consistency in applying codes. Their expectation is that consistency in coding provides proof that the identified concepts would be the same even if a different researcher were to apply the same codebook to the same data. Consistent coding is especially important when conducting research that can critically affect individuals, such as deciding meaningful endpoints in clinical trials (Krippendorff, 2022). Yet the definitions of coding agreement are often muddled, and the steps of achieving agreement have not been well defined. I address this issue by providing an overview of various inter-coder agreement (ICA) tests and provide a framework for establishing a robust ICA process. In the current manuscript I will provide guidance that can help researchers determine whether statistical methods of coding comparison are appropriate for their study, and if so, steps for applying them in a systematic way.

[1] Corresponding author; is a Qualitative Analyst with RTI-International. 3040 E Cornwallis Rd. Research Triangle Park, NC 27709, USA. Ph: 770-234-5047. E-mail; snhalpin@rti.org

Trustworthiness refers to "quality, authenticity, and truthfulness of findings of qualitative research" and is often used synonymously with reliability and replicability (Cypress, 2017, p. 254). Statistical methods of coding comparison are appropriate when aiming for coding replicability, even among coders who are not involved in developing the codebook, ensuring external reliability. External reliability is important when concerned with research that directly impacts individuals, such as setting outcome objectives in clinical trials based on qualitative interviews of patients with a particular health condition (Krippendorff, 2022). Alternatively, researchers may use statistical methods of coding comparison to demonstrate internal reliability (only replicable among the team of coders involved in creating and revising the codebook). As one group of researchers put it, "we were concerned with the dependability of the coders working on the same team not the replicability of the instrument across different teams or projects" (Cascio et al., 2019, p. 4). Internal reliability is desirable when training someone new to qualitative coding or when engaging multiple coders in a single project. Nevertheless, I agree with other researchers that using statistical methods of coding comparison when only seeking internal reliability and not external reliability is inappropriate and risks creating the illusion of rigor (Cook, 2012; Krippendorff, 2022; J. Morse, 2020; J. M. Morse, 1997, 2015; Yardley, 2000). Furthermore, using statistical methods of coding comparison only for internal reliability may lead researchers to forgo other, more appropriate methods of assuring trustworthiness such as saturation (Francis et al., 2010; Guest et al., 2006), clearly stating the data collection process (O'Sullivan & Jefferson, 2020), whether the sample is adequate (Sim et al., 2018), reflexivity (Watt, 2015), thick description (Geertz, 1973/2021), negative case analysis (Denzin, 2017), multiple forms of triangulation (i.e., method, investigator, theory, and data source) (Denzin, 2017), and member checking (Lincoln et al., 1985). This is a non-exhaustive list of trustworthiness methods, and importantly, researchers should demonstrate trustworthiness aligning with their chosen methods. For example, saturation is not appropriate for reflexive thematic analysis (Braun & Clarke, 2021a.). Nevertheless, researchers should be intentional regardless of the method they use to demonstrate the trustworthiness of their qualitative study and, when appropriate, should seek to establish external reliability.

Detractors of mathematical formulas for coding comparison often claim that establishing objective and impartial coder agreement is unachievable, particularly in studies using inductive data (i.e., using Inter-Rater Reliability (IRR); Cook, 2012; Krippendorff, 2022; J. M. Morse, 1997, 2015, 2020; Yardley, 2000), given these studies benefit from varied coder perspectives and result in investigator triangulation (Denzin, 2017). In these cases, it is considered beneficial to have multiple investigators, each making independent observations, who then come together and compare results. Researchers espousing this view aim to examine questions from a subjectivist epistemology, emphasizing how each individuals' independent experiences influence how they view and interpret the external world (Crotty et al., 2020). Alternately, the use of deductive coding and ICA as a method for demonstrating coder agreement may be well suited to positivist-aligned methods, given their belief in pre-existing and objective knowledge (Cook, 2012; Díaz et al., 2023; McDonald et al., 2019; Nili et al., 2020). An example of a theoretical approach aligning with deductive coding is critical realism, where

> Reality is stratified into three domains: the domain of the 'real' (made up of these natural and social objects, structures and their mechanisms) the 'actual' (comprised of events, that is, what happens when mechanisms are activated) and the 'empirical' (which refers to our perceptions and experiences of these events) (Hoddy, 2019, p. 112).

Even when the participants' epistemological assumptions align well with using coding comparison measures analysis utilizing multiple coders may still suffer from the possibility that coders may re-enforce their own biases (Krippendorff, 2022; Moret et al., 2007). For instance, two

coders applying the code, finite cognitive processing, which is not well defined in the codebook, may rely on their own background to help interpret how the code should be applied (Halpin & Konomos, 2022; Halpin, Konomos, & Jowers, 2021). Regarding Inter-Coder Agreement (ICA) and deductive coding, a solution may include well defined codes, limiting the likelihood for pre-defined prejudice influencing coding. Another pitfall includes the possibility that,

> One can imagine that at a certain point, they [coders] no longer judge on the grounds of a 'mature judging skill' but do so as 'brainwashed' automaton reacting to stimuli and accurately assigning codes in the way they expect the researcher [qualitative data manager] would like them to (Muskens, 1980, p. 124).

Yet here again, the concern regarding "brainwashed" coding may apply more often to inductive studies using IRR, where the researchers' varied experiences are considered to add analytical depth. Finally, once coding disagreement is identified, researchers often state that they met to resolve any discrepancies, which may lead to a tit-for-tat reconciliation where one coder may give in and allow another coder to make the decision or else allow the researcher who feels more passionate about a code or holds a more senior role, to make the final decision (Clarke et al., 2023).

Multiple mathematical formulas exist for verifying harmony between two or more researchers' coding either using a pre-established, deductive codebook (termed Inter-Coder Agreement [ICA]) or when applying new emergent or inductive codes (termed Inter-Rater Reliability [IRR]). Others differentiate Inter-Coder Reliability as categorized at a nominal level (e.g., binary categories) versus IRR coding on an ordinal or interval level (e.g., to a greater or lesser degree) (O'Connor & Joffe, 2020). These researchers further define Intra-Coder Reliability as consistency in an individual's coding over time (Joffe & Yardley, 2004; Kurasaki, 2000). In any case, methods of coding agreement may help assure a health science audience of trustworthiness in coding. Yet, qualitative researchers have not always embraced coding agreement as a measure of trustworthiness but may be required to do so depending on the beliefs of research team members, other project requirements, or their intended publication outlet.

Many books and articles exist outlining methods for using IRR and ICA, but each lacks in some regard, such as having a focus on inductive studies (Braun & Clarke, 2006, 2021; Campbell et al., 2013; Charmaz, 2006; Cole, 2023; Compton et al., 2012; O'Connor & Joffe, 2020; Roberts et al., 2019) and therefore do not address pitfalls discussed above. For example, one recent guidance publication did not provide advice on how to work through coding disagreements and how those decisions should be documented and presented in publications (O'Connor & Joffe, 2020).

The current manuscript focuses on ICA, given that the approach is a better fit for deductive analysis and seeking external validity. Deductive qualitative analysis may be more often used when studies rely on pre-existing theory and/or pre-existing literature. Codes can also be initially developed inductively; however, in the current manuscript, I will start with the requirement that ICA is conducted once no additional codes will be added. The reason for this distinction is, as mentioned above, the need for coding that is replicable not only within a small group of individuals who developed the codebook—but rather is replicable regardless of who applies the codes. As deductive qualitative methods are used increasingly in health sciences and in various other fields, understanding how to conduct and interpret ICA is important. In this manuscript, I attempt to address these gaps and provide a thorough guide to conducting ICA, including an outline of the essential steps and methodologies. Through exploration of ICA techniques, I seek to demystify ICA and make it more accessible and intentional.

Planning for ICA

Researchers who choose to use ICA should proactively consider the test they plan to use, their threshold of acceptability, the method of calculation, what data will be double coded, codebook development, and how many coders will code the data. Also, researchers should create a plan for addressing any codes not meeting their established threshold of acceptability and for tracking any changes to the codebook. Study teams should assign a Qualitative Data Manager at the planning stage, who is responsible for tracking decisions and for steps such as distributing coding assignments, merging coded data, and running ICA calculations.

Determine Which Test to Use

Many mathematical tests of ICA exist, each with subtle differences. These methods may be conducted with or without QDAS. If conducted within QDAS, it will be important to consult the program you are using to determine which ICA tests are available for analysis within the program. Below are the most reported ICA tests, but it is also important to recognize that other tests exist, including Scott's Pi (Scott, 1955) and Fleiss' kappa (Fleiss, 1971). Additionally, variants exist on the methods for calculating ICA, some of which may not be available in different QDAS programs. No single test of ICA is perfect. The best test depends on the research question. Researchers should carefully consider the strengths and weaknesses of each test before choosing which one to use (Table 1).

Table 1

When Different ICA Tests are Appropriate

Test	Disadvantages	When to use
Percent agreement	• Influenced by number of codes in codebook and likelihood of those codes being applied by chance.	• The codebook is straight forward and simple. • The coders are experienced and have a good understanding of the codebook.
Holsti's method	• Provides an overall agreement score- rather than pointing to specific codes where agreement is not met.	• The codebook is longer and less straight forward.
Cohen's kappa *(k)*	• Possibility of discordance between high agreement and a low kappa value. • Cases where some codes are used more frequently than others in the codebook can result in overestimation of observed agreement. • Can be more easily influenced by smaller sample sizes.	• Your codes are applied evenly across the dataset. • You have a larger sample size.
Krippendorff's alpha	• Cases where some codes are used more frequently than others in the codebook can result in overestimation of observed agreement. • Challenging to calculate by hand.	• You have more than two coders. • You have a larger and more complex dataset.
Gwet's AC1 and AC2	• You have a smaller sample	• You have more than two coders. • You want to account for total agreement expected by chance. • You have a larger and more complex dataset.

Percent Agreement. Percent agreement constitutes the simplest ICA measure (Miles et al., 1994). The measure is calculated as the number of times two or more coders agree on a given code divided by the total number of codes and multiplied by 100, with higher numbers representing higher agreement. However, percent agreement is influenced by how many codes exist in the codebook and the likelihood of codes being applied by chance (Krippendorff, 2022). As such, a set of coders could have a different understanding of the same text but still report a high percent agreement if they agreed on the same number of codes. Using percent agreement might be best under the following conditions: (1) the codebook is straightforward and simple, and (2) the coders are experienced and have a good understanding of the codebook.

Holsti's Method. Holsti's method builds on Percent Agreement by measuring the aligning in text that is coded between researchers (Parker & Holsti, 1970). The calculations can be used for two coders and is calculated by two multiplied by the number of agreements and divided by the sum of the number of spaces coded (i.e., the letters, punctuation, and spaces between words) by the two researchers. As such, Holsti's method is better suited for cases where two coders have not

applied the exact same number of codes. Possible scores range from 0 to 1, with higher scores indicating higher reliability. While Holsti's method is less likely to inflate chance agreement, it ultimately provides an overall agreement number, which does not allow for identifying codes with lower rates of agreement. Using Holsti's method might be best under the following conditions: (1) the codebook is longer, and (2) less straight forward.

Cohen's Kappa (k). Cohen's kappa involves two coders and accounts for chance agreement when establishing ICA (Cohen, 1960, 1968). Scores ranging from 0 to 1, with higher scores indicating higher agreement. The test is calculated by subtracting the expected agreement from the observed agreement, and then dividing the number by one minus the expected agreement. While Cohen's kappa is the most commonly reported ICA test, the results can be misleading, including achieving a low Cohen's kappa despite high agreement (Krippendorff, 2004; Xie, 2013). A second problem includes unbalanced marginal distributions producing higher kappa values than balanced marginal distributions (Xie, 2013). Marginal distribution refers to the total number of times each coder assigns a particular code, and the number is used to calculate the expected level of agreement (the number of times the two coders would be expected to agree on a code by chance). If a code is disproportionately relevant in a dataset compared to other codes, the observed agreement risks being overestimated. In these cases, a weighted kappa may be used to offset the differences. Using Cohen's kappa might be best under the following conditions: (1) your codes are applied evenly across the dataset, and (2) you have a larger sample size.

Krippendorff's Alpha (a). Krippendorff's alpha may be used with more than two coders at a time and accounts for chance agreement (Krippendorff, 2022). Further, the measure accounts for the number of codes used and the prevalence of codes to combat a major criticism of Cohen's kappa (Xie, 2013). As such, Krippendorff's alpha may be particularly useful for assessing the reliability of coding in large datasets and for coding complex data. Results range from 0 to 1, with higher scores indicating higher agreement. Nevertheless, the measure suffers from some of the same drawbacks as Cohen's kappa, namely, it is not always clear if a high observed agreement is due to reliability or high chance agreement. Further, the calculation has been criticized for being complex and thus difficult to calculate by hand (Lombard et al., 2002). Yet, authors have attempted to develop strategies for simplifying the calculation of Krippendorf's alpha (González-Prieto et al., 2023). Using Krippendorff's alpha might be best under the following conditions: (1) you have more than two coders, and (2) you have a larger and more complex dataset.

Gwet's AC1 and AC2. Gwet's AC1 and AC2 are each chance-corrected measures, meaning they account for agreement that would be expected by chance (Gwet, 2010). Gwet's AC1 is a measure of overall agreement between two or more coders and is calculated by subtracting the expected chance agreement from the observed agreement and then dividing by the maximum possible agreement. Meanwhile, Gwet's AC2 is a measure of the agreement between two or more coders on the presence or absence of a particular category. Gwet's AC2 is calculated by subtracting the expected chance agreement from the observed agreement and then dividing by the sum of the observed agreement and observed disagreement. Both Gwet's AC1 and AC2 are reported to have a possible range of 0 to 1, with higher scores indicating higher agreement. Gwet's AC1 was found to be more reliable than Cohen's kappa in a sample of patients with personality disorders, especially in cases where the prevalence of the disorder in the study population was low and, therefore, those codes were used less frequently (Wongpakaran et al., 2013). Still, the measures may be sensitive to cases when the distribution of codes is uneven. Using Gwet's AC1 and AC2 might be best under

the following conditions: (1) you have more than two coders, (2) you want to account for total agreement expected by chance, and (3) you have a larger and more complex dataset.

Determine the Threshold of Acceptability

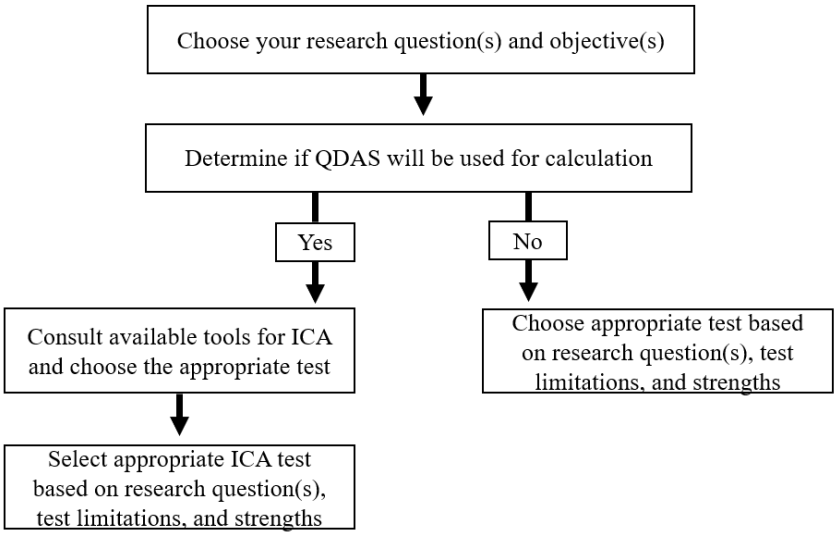
Threshold of acceptability is often presented as a range from 0 (no concordance) to 1 (full agreement), with numbers often converted to percentages. Some guidelines recommend a minimum 0.8-0.9 ICA threshold (De Munck, 2000; Lombard et al., 2002; Saldaña, 2016). Meanwhile, Miles et al. (1994) have recommended an 0.8 threshold on 95% of the codes—meaning 5% of codes may be allowed to fall below the 0.8 agreement threshold and still be considered reliable. Alternately, Landis and Koch (1977) recommended a gradient where 0.41-0.6= moderate, 0.61-0.8= substantial, and 0.81-1= almost perfect reliability. Alternately, Krippendorff (2022) has recommended that an ICA of 0.6 may simply occur due to chance coding. The appropriate threshold for ICA will vary depending on several factors defined by the researchers, including the purpose of the research and the desired level of confidence. As such, higher thresholds are related to higher rates of agreement. For example, in studies researchers seeking to confirm earlier agreed upon criteria who desire a high level of confidence, as defined by the researcher, such as when defining meaningfulness changes in symptoms for future clinical trials, may select a higher threshold. However, researchers who are willing to accept a lower level of confidence may choose a lower threshold, such as cases with more complex codebooks and smaller samples.

Determine Your Method of Calculation

Multiple methods exist for calculating ICA, including by hand, QDAS, and importing data to a statistical analysis software or online coding program. Below is a brief explanation including advantages and disadvantages one should consider when determining your method of ICA calculation. Researchers should carefully consider what data analysis platforms are available to them, whether they have the knowledge and ability to apply their calculation methods for those platforms, and as stated above, which test is most appropriate for their objective (Figure 1).

Figure 1

Decision Flowchart for Selecting the Appropriate Intercoder Agreement (ICA) Test Based on Research Goals and Data Analysis Platforms



ICA Calculated by Hand. Calculating ICA without the aid of computer software can be cumbersome and prone to error, especially when considering more complicated statistical tests (McAlister et al., 2017). Percent agreement is the most reasonable test to calculate by hand, but as mentioned above, in many cases this test is not ideal.

ICA Calculated by QDAS. Qualitative data analysis software (QDAS) offers many benefits for researchers using ICA (Woods et al., 2016; Zamawe, 2015). The available QDAS tools enable coders to efficiently code and organize qualitative data, which makes tracking coding decisions and ICA discrepancies easier. The software can also streamline the process of quantifying ICA through built-in statistical calculations. Moreover, the software can increase transparency by documenting the coding process, including tracking changes to the codebook through an audit trail, helping researchers establish the reliability of their analysis. These advantages may be desirable, especially in cases where there is an expectation for providing a quantitative comparison of coding, such as from a funder or target journal. However, disadvantages exist, especially in cases of inductive coding, including the reduced coder autonomy and potentially missing nuanced insights. Further, the cost and learning curve associated with using QDAS may be a barrier. More broadly, different QDAS and their different versions offer varying types of built-in statistical calculations. For example, Krippendorff's alpha is not present within NVIVO version 12.0, though it does offer percent agreement and Cohen's kappa. Similarly, ATLAS.ti version 23 does not offer Cohen's kappa, though it does offer Krippendorff's alpha.

ICA Calculated by Statistical Software. Statistical software suites, such as Statistical Analysis System (SAS) and others, may also be used for calculating ICA in a similar way to QDAS, though the user will not be limited by the statistical test that is native to the program. Using these programs, researchers can calculate frequencies of applied codes and apply the chosen statistical test. Importantly, calculating ICA using statistical software suites is not as user-friendly as more intuitive QDAS and their process may require a steeper learning curve along with statistical expertise.

ICA Calculated by Programing Language. Lastly, researchers could use an online coding and analysis library, such as the Python library '**codinganalysis'** (Marzi et al., 2024). Much like statistical software suites, described above, programing languages allow researchers to load their qualitative data using various formats, and calculate the ICA test of their choosing. However, also like statistical software suites, these methods require programing knowledge and careful attention to methodological rigor.

What Data Should be Coded by More than One Coder (Double Code)

The amount of data to double code for ICA depends on varied factors, including the complexity of the research, the number of coders, the research objectives, and funder preferences. Researchers should begin with a subset of the dataset that will give enough data to assess consistency in coding across the codebook, but not so much that excessive time is invested only to discover inconsistencies that could have been rectified sooner. Previous guidance has recommended to begin by double coding the first transcript (O'Connor & Joffe, 2020), while others have provided less specific advice and instead encouraged researchers to, at a minimum, code the entire transcript rather than an excerpt (Feeley & Gottlieb, 1998; Graneheim & Lundman, 2004). Researchers must balance their budget, time resources, and coder expertise when deciding how

much of the data should be double coded. With more complex data, and/or in cases where coders are less experienced in the content area, it is best to double code a larger proportion of data. In cases where only a subset of data is double coded, researchers should periodically review the data to combat the chance of coders changing how they approach the coding over time, as discussed in greater depth later in this manuscript.

All codes in a codebook should be applied with enough frequency in order to calculate ICA. For example, if ICA is calculated using a dataset where a portion of codes within the codebook are not applied, it will be impossible to assure consistency in the application of those non-applied codes. Further, as mentioned above, some statistical tests are sensitive to whether codes are applied more often and others infrequently within the collection of data used for ICA.

Codebook Development

The overall objective of ICA should be to produce a codebook that can be used by any qualitative researcher to achieve a high degree of reliability without further changes. As such, the goal of ICA should not be to develop a codebook which is only valid within the individuals who created that codebook. Thus, developing an initial deductive qualitative codebook is a crucial step in the ICA process (Roberts et al., 2019). To create a deductive qualitative codebook, researchers should begin by thoroughly reviewing published literature and established theories on concepts relevant to the research aims. The key themes and constructs identified from literature and theories will serve as the foundation for the initial codes. Next, the researcher should create a list of codes from the previously identified themes along with clear definitions. Researchers should thoughtfully map codes for each study aim, with consideration for how the results will be presented at the end of the project. Many of the methods for calculating ICA are sensitive to variation in how often the codes are used, and therefore it is critical to consider whether each code will contribute to the study objectives. Therefore, it is best to include codes that will contribute to the overall objective. Additionally, it is not reasonable to believe that coders will be able to accurately recall an excessive number of complex codes. MacQueen and colleagues (2008), for example, recommended that coders should only be expected to apply 30-40 codes in one sitting. Therefore, it might be a good guide to limit the number of codes, and in cases with a higher number of codes, the coding process should be approached in stages. An excerpt example codebook, created for the current manuscript and based on a bi-nary theoretical framework, is provided in Table 2 (Halpin, Dillard, and Puentes, 2017).

Table 2

Deductive Codebook Excerpt Example

Code Name	Code Definition
5. Coping	Code all text that refers to how a person is dealing with cognitive impairment. The focus here is on the patient but this code can also be applied to caregivers. “5. Coping”
5.1 Adaptive	Refers adjusting to changing functional status related to cognitive impairment. (Ex; Using a calendar; using an alarm reminder to take medications)
5.2 Non-adaptive	Refers to any instance where the patient does not adjust to their changing functional status related to cognitive impairment.

How Many Coders?

The number of coders in ICA depends on the research goals, available resources, and the complexity of the coding. While at least two coders are necessary to achieve ICA, it may be desirable to have three to five coders to enhance the permutations of the comparisons between coders or based on timelines and availability of coders (Devotta & Pedersen, 2015). In cases where more than two coders are used, it can be helpful to stagger pairs of coders, so the pairs change throughout the ICA process (Devotta & Pedersen, 2015). For example, coder A and B may code the first transcript, and then coders B and C code the second, followed by coders C and A, and so on. The advantage of a staggered approach includes assurance that the entire team is coding in a similar pattern. However, involving too many coders can introduce challenges related to consistency and coordination. The right balance is essential. Others have recommended that at least one coder should have expertise in the content area and previous experience coding (Cofie et al., 2022). Though, in cases where there exists an imbalance in content knowledge, it is important to carefully consider in advance how coding disagreements will be resolved, as discussed in the next section (Cheung & Tai, 2023). Ultimately, the choice of the number of coders should be a well-considered decision, aiming to strike a balance between comprehensive analysis and the practical feasibility of managing the coding process effectively.

Develop a Process for Working through Disagreements

Resolving differences in qualitative coding is an essential part of ICA and a topic that is often overlooked and/or not fully reported in published literature. The process itself is often described in a single line, such as 'the researchers met and resolved disagreements,' with little detail about how the disagreements were resolved (O'Connor & Joffe, 2020). Others have more thoroughly laid out processes where two coders should systematically review any areas where they disagree, argue their point, and then decide (Allsop et al., 2022). However, the process of working through disagreements is often one of the most turned-to critiques of why some consider ICA inappropriate. As Krippendorf (2004) stated, "Resolving disagreements by majority among three or more coders may make researchers feel better about their data but does not affect the measured reliability" (p. 219). It is assumed that a negation process is employed, where two or more coders reach an agreement through discussion and likely compromise. Detractors suggest that the compromise element can be problematic because coders may compromise on coding criteria due to a power imbalance or a 'tit for tat' concession where researchers feel indebted to the other coder(s) or beholden to move the coding forward for timeline reasons, and will agree to coding changes they otherwise would not (Clarke et al., 2023; J. M. Morse, 1997; Zade et al., 2018).

Non-negotiation strategies exist to work through disagreements. Zade and colleagues (2018), for example, suggest that a neutral third party should be responsible for mediating or settling disagreements between coders. While it may not be possible for a third party to be completely neutral, strategies exist for increasing transparency and impartiality. Armstrong and colleagues (1997) suggests that rather than a completely neutral third party, a principal researcher or an independent panel could be used to settle disagreements. Below are important considerations when outlining a process for working through disagreements based on conflict resolution theory (Bransford et al., 1998). Importantly, this process should be facilitated by the Qualitative Data Manager, who should function as an independent third party. The third party would need to have expertise in both the content area as well as qualitative analysis to adequately address differences in a way that minimizes the likelihood of simply reinforcing their own biases.

1. Identify the problem: The identified problem in this case will be each code that does not meet the pre-established threshold of acceptability.
2. Gather information: The Qualitative Data Manager may request coders provide explanations for coding differences that the third party can then review. Additionally, the Qualitative Data Manager should review the coding discrepancies, identifying what text each coder applied the codes to.
3. Generate solutions: After the reviewing the gathered information, the Qualitative Data Manager should choose a solution, with the aim of developing a coding structure that would be reliably applied both by the current coders and any other coders not originally involved in the coding process.
4. Implement the solutions: The solution may involve updating definitions of the codebook based on the decision chosen in step three, and/or re-education of the coder(s) based on the decision. The revised coding files and updated codebook should be managed and distributed by the Qualitative Data Manager.

Create an Audit Trail Plan

Creating an audit trail is a fundamental practice for tracking changes made to a qualitative codebook, promoting transparency and accountability in the research process, and providing documentation that can be referred to in presented data. The audit trail process involves documenting each modification, revision, or addition to the codebook over time, including details on the date of change, description of the change, and rationale for the change (Miles et al., 1994). An example of an audit trail for codebook changes, based on a framework described above and made for illustrative purposes in this manuscript, can be found in Table 3. Often, QDAS maintains an audit trail, but the process can also be maintained by a member of the research team using a spreadsheet. In either case, the researchers should plan an audit trail strategy early and be consistent and specific about what changes were made.

Table 3

Audit Trail Example

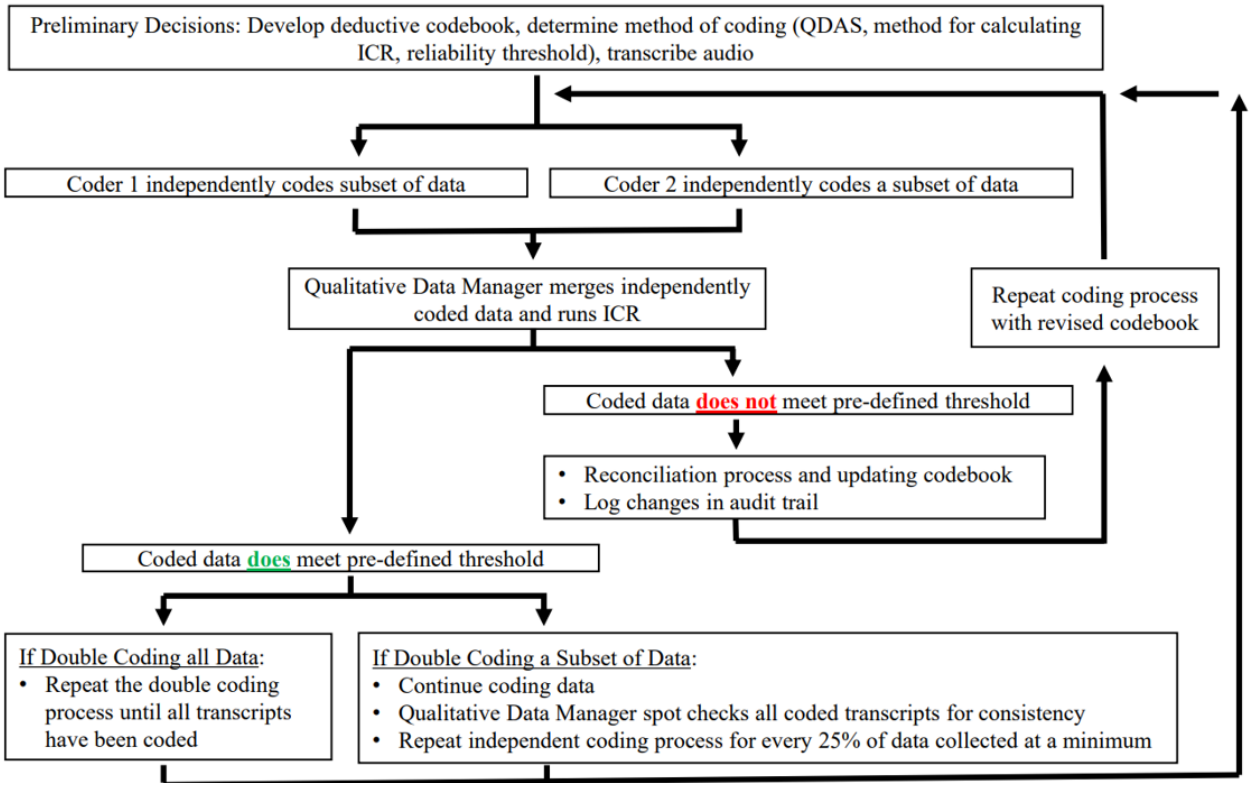
Date of Change	Description of the Change	Rationale for the Change
10/21/2015	Code 5.1 Adaptive was updated to include the following text “Ex; Using a calendar; using an alarm reminder to take medications”	Change made to provide examples of when code 5.1 Adaptive should be applied.

Applying ICA

Researchers must carefully consider the steps necessary for applying ICA, along with when they should return to the documents created during planning such as the process for working through disagreements and audit trail. This section elaborates on ICA coding and reconciliation process outlined in Figure 2. While the figure envisions two independent coders, it is possible to have additional coders as defined in the planning phase above.

Figure 2

ICA Coding and Reconciliation Process



Step 1. Independently Coding a Subset of Data

Coders should independently code the same set of data using the same codebook. The data to be used for the initial coding is chosen in the planning phase (described above) when choosing what to double code. Importantly, this should be a representative amount of data that will provide a good opportunity to use all codes in the codebook. It may be beneficial to code an entire interview transcript, for example, to give complete context for a particular sitting (Feeley & Gottlieb, 1998; Graneheim & Lundman, 2004; Roulston & Halpin, 2022).

The coders should begin by familiarizing themselves with the codebook and subset of data they will be coding. The coders should keep track of any questions that arise when applying codes, such as cases when they are unsure about whether to apply a code and the reasoning for the confusion.

Step 2. Qualitative Data Manager Merges Independently Coded Data and Runs ICA

Step two should be completed by a Qualitative Data Manager, who was not one of the two independent coders from step one. The reason for the Qualitative Data Manager not being one of the two independent coders is to decrease the level of subjectivity in the process. The Qualitative Data Manager should follow the steps agreed upon in the preliminary planning stages, including whether QDAS will be used for running ICA.

Step 3. If Coded Data Does Not Meet Pre-Defined Threshold

If the completed ICA does not meet the pre-determined threshold, researchers will need to follow the following steps, first completing a reconciliation process and logging those decisions in the audit trail, and then repeating the coding process with the revised codebook.

Step 3a. Reconciliation Process and Audit Trail. The Qualitative Data Manager should be responsible for completing the reconciliation process in cases where agreement does not meet the pre-established threshold. Additionally, decisions should be updated in the codebook and audit trail.

Step 3b. Repeat Coding Process with Revised Codebook. After the reconciliation process and audit trail have been logged by the Qualitative Data Manager, the independent coders should return to the data they coded previously and re-code based on the revised codebook. It is acceptable at this point for coders to return only to excerpts where the revised codes are applicable if that can be determined. Once done, the Qualitative Data Manager will repeat step two and proceed according to whether the ICA threshold is met or not. The process should be repeated until the pre-established ICA threshold is met. Please note, others have argued that the original coders should not return to their previously coded data, and rather a new second coder who has not yet worked with the data should apply the refined codebook (Krippendorff, 2022).

Step 4. If Coded Data Does Meet Pre-Defined Threshold

If the completed ICA does meet the pre-defined threshold, the researchers will continue double coding the agreed-upon subset of data. The Qualitative Data Manager should plan to repeat the process in Figure 2. at a minimum every 25% of completed data. Further, the Qualitative Data Manager should randomly review at least 10% of the remaining coded data to ensure coding consistency. This process will help guard against the possibility of coding creep or inconsistency in how coders apply the codebook over time (Belur et al., 2021; Rousson et al., 2002; Halpin, 2023).

Reporting

Reporting the process used to achieve ICA helps improve the transparency and reproducibility of results. Increased transparency may help ensure the quality and rigor of the research process. Researchers should focus on three categories: (1) organization, (2) clarity, (3) conciseness, completeness, and accuracy, and in communicating the results to the reader (J. L. Johnson et al., 2020). The text below builds on previous reporting guidance and includes direction on how to effectively report the ICA process considering each of these three categories, including when faced with word limit constraints such as in journal articles (Prasanth, 2021).

Organization

Several qualitative study checklists exist, though they either do not include ICA or only refer to it briefly. For example, the Consolidated criteria for Reporting Qualitative Research (COREQ) found that studies sometimes report on inter-observer reliability but do not include coder comparison in their checklist (Tong et al., 2007). Similarly, the Critical Appraisal Skills Programme (CASP) has a single sentence stating that reliability is often used for positivists research (Long et al., 2020). Finally, both the Rigour and Transparency in Qualitative Research

(RATS) and the Standards of Reporting Qualitative Research (SPQR), lack guidance on using ICA (Mays & Pope, 1995; O'Brien et al., 2014).

Whether in a manuscript or another publication method, the product should be organized to provide a comprehensive and well-structured overview, guiding readers systematically through the ICA process in a clear chronological order. The organization should outline the specific ICA test employed and elucidate the criteria and parameters used to establish the threshold of acceptability and how the test was conducted (e.g., using QDAS). Additionally, the organization should elaborate on the number of coders actively engaged in the ICA process, detailing their qualifications and experience in both qualitative research and the content area examined. Finally, the organization should include space to discuss how potential disagreements arising during coding were addressed and tracked.

Clarity

Clarity in academic writing is the foundation of effective communication and knowledge dissemination. Clarity supports the transmission of complex information and concepts to a wider audience, making the scholarly discourse accessible and comprehensive. As with all academic writing, researchers tasked with reporting ICA findings must strive to achieve a balance between being accessible and comprehensive. Researchers should articulate the purpose of their work while avoiding the pitfalls of excessive reliance on lengthy lists of prepositions and verbs (Sword, 2015, 2018). The goal is to present information in a manner that is clear, concise, and engaging.

Conciseness, Completeness, and Accuracy

Whether faced with word limit constraints for a journal article or endless digital space on a blog posting, researchers should provide a concise, complete, and accurate accounting of the ICA process. Despite the challenge of adhering to word limits, researchers should present a complete and accurate accounting of their ICA process. When possible, these details should be reported in the text, but they may also be reported using supplemental documents if the journal allows. Alternately, in platforms that afford greater writing space, such as blogs or book chapters, researchers should maintain a commitment to concise writing. In all cases, the objective remains the same: to uphold the integrity of the research by ensuring the account is not just succinct but also comprehensive and accurate.

Discussion and Conclusions

The current manuscript includes detailed guidance for ICA use in deductive qualitative studies. ICA, for qualitative research, is an imperfect tool. Indeed, many non-quantitative methods exist for demonstrating trustworthiness in qualitative research, including saturation (Francis et al., 2010; Guest et al., 2006), clearly stating the data collection process (O'Sullivan & Jefferson, 2020), whether the sample is adequate (Sim et al., 2018), reflexivity (Watt, 2015), thick description (Geertz, 1973/2021), negative case analysis (Denzin, 2017), triangulation (Denzin, 2017), and member checking (Lincoln et al., 1985). Nevertheless, in cases where researchers feel compelled to use ICA, it is critical to carefully consider in what cases ICA is most appropriate and the potential pitfalls when using ICA. Much guidance for performing ICA has covered the gamut of qualitative research, including disparate ontological stances (i.e., whether reality exists) and epistemological stances (i.e., whether reality is discoverable) (Braun & Clarke, 2006, 2021b.; Charmaz, 2006; Cole, 2023; O'Connor & Joffe, 2020; Roberts et al., 2019).

In this article, I have argued that ICA is most appropriate in situations when coders are being asked to apply deductive, or previously defined codes, as deductive coding better aligns with the positivist nature of ICA. Moreover, researchers should carefully consider in what context they want their coding to be reliably applied. I have argued that ICA is best suited for cases when external reliability is desirable, such as when results of the study will be used to decide on meaningful endpoints in clinical trials. Alternatively, in cases where only internal reliability is sought, ICA may not be the most appropriate method for demonstrating trustworthiness. Lastly, studies benefit from having multiple forms of trustworthiness regardless of whether their objective is to achieve external or internal reliability. As such, applying multiple additional forms of non-statistical methods of demonstrating trustworthiness, such as those referenced above, should be applied with equal care.

The current study also advances on the theoretical understanding of ICA by examining the complexities of coding discrepancies and proposing systematic approaches for their resolution. Further, by drawing on conflict management theory, this manuscript sheds light on the dynamics of the collaborative coding process and offers insights into effective strategies for achieving consensus among coders. Additionally, by critically examining existing practices and identifying areas for improvement, this manuscript lays the groundwork for enhancing empirical validation and theory building in qualitative research methodology.

Despite best efforts to provide a clear and concise approach to using ICA, there are limitations to the proposed approach. Notably, the above approach is resource-intensive in terms of time, expertise, and finances, which could make ICA prohibitive. I argue that if research teams are unable to follow this rigorous approach, they should instead focus on other methods of trustworthiness, as described above. Additionally, the process of achieving ICA does not guarantee the generalizability and validity of a researcher's findings, qualities that are rarely achievable in qualitative studies (Leung, 2015). As such, other factors, such as the sampling strategy, depth of the analysis, and integration of multiple sources of evidence, should be considered.

Future studies may consider how to best integrate ICA into mixed-methods studies using qualitative and quantitative data. For example, the process described above may need to be modified to consider when different types of data are interrogated simultaneously (R. B. Johnson & Onwuegbuzie, 2007). Using a simultaneous approach would necessitate the quantitative data influencing further exploration of the qualitative data and, as such, would likely involve nuanced changes to the codebook, which could further complicate the ICA process. Secondly, researchers may explore the optimal length and complexity of a codebook to increase the likelihood that coders will retain their nuanced understanding of the code definitions and apply the codes consistently over time. As discussed above, researchers have made suggestions on the total number of codes that should be included in a codebook, yet these ideas are not empirically validated (De Munck, 2000). Future studies could track the ICA process as it relates to the numbers of complexity of codes, and perhaps make recommendations on what factors lead to more productive coding, as relates to factors such as time taken to code. Lastly, future research could consider optimal thresholds of acceptability for different ICA tests based on metrics such as how many rounds of revision it takes to reach certain thresholds.

Researchers who do choose to use ICA should carefully consider their plans for approaching and tracking the ICA process before engaging in the practice, as outlined above. In particular, researchers should consider which statistical test they will use, their threshold of acceptability, the method of calculation, which data to double code, their codebook development, how many coders will code the data, and a plan for addressing any codes that do not meet the threshold. Roles for coders and a Qualitative Data Manager should be assigned upfront as well.

Further, carefully considering the process of resolving any coding discrepancies, as describe in the steps above, along with how that process will be recorded and reported as critical steps for ensuring ICA helps qualitative researchers assure readers of the trustworthiness of their data.

Acknowledgement

Thank you to Sara Andrews, Megan Lewis, Alison Halpin, and Abigail Halpin for your invaluable input in editing this manuscript.

Funding Statement

Preparation of this manuscript was supported by the RTI International Fellow Program.

Conflict Of Interest Statement

The author reports no conflicts of interest.

References

Allsop, D. B., Chelladurai, J. M., Kimball, E. R., Marks, L. D., & Hendricks, J. J. (2022). Qualitative methods with Nvivo software: A practical guide for analyzing qualitative data. *Psych*, *4*(2), 142–159. https://doi.org/10.3390/psych4020013

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: An empirical study. *Sociology*, *31*(3), 597–606. https://doi.org/10.1177/0038038597031003015

Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods and Research*, *50*(2), 837–865. https://doi.org/10.1177/0049124118799372

Bransford, J. D., Haynes, A. F., Stein, B. S., & Lin, X. (1998). *The IDEAL workplace: Strategies for improving learning, problem solving, and creativity*. NashvilleREAD.Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2021a). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–252. https://doi.org/10.1080/14780887.2020.1769238

Braun, V., & Clarke, V. (2021b). To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative Research in Sport, Exercise and Health*, *13*(2), 201–216. https://doi.org/10.1080/2159676X.2019.1704846

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semi-structured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods and Research*, *42*(3), 294–320. https://doi.org/10.1177/0049124113500475

Cascio, M. A., Lee, E., Vaudrin, N., & Freedman, D. A. (2019). A team-based approach to open coding: Considerations for creating intercoder consensus. *Field Methods*, *31*(2), 116–130. https://doi.org/10.1177/1525822X19838237

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis (Introducing Qualitative Methods series)*. SAGE Publications.

Cheung, K. K. C., & Tai, K. W. H. (2023). The use of intercoder reliability in qualitative interview data analysis in science education. *Research in Science and Technological Education*, *41*(3), 1155–1175. https://doi.org/10.1080/02635143.2021.1993179

Clarke, S. N., Sushil, S., Dennis, K., Lee, U. S., Gomoll, A., & Gates, Z. (2023). Developing shared ways of seeing data: The perils and possibilities of achieving intercoder agreement. *International Journal of Qualitative Methods*, *22*(1), 1–10. https://doi.org/10.1177/16094069231160973

Cofie, N., Braund, H., & Dalgarno, N. (2022). Eight ways to get a grip on intercoder reliability using qualitative-based measures. *Canadian Medical Education Journal*, *13*(2), 73–76. https://doi.org/10.36834/cmej.72504

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

Cole, R. (2023). Inter-rater reliability methods in qualitative case study research. *Sociological Methods and Research.*. Advance online publication. https://doi.org/10.1177/00491241231156971

Compton, D., Love, T. P., & Sell, J. (2012). Developing and assessing intercoder reliability in studies of group interaction. *Sociological Methodology*, *42*(1), 348–364. https://doi.org/10.1177/0081175012444860

Cook, K. E. (2012). Reliability assessments in qualitative health promotion research. In *Health Promotion International*, *27*(1), 90–101. https://doi.org/10.1093/heapro/dar027

Creswell, J. W. (2003). *Research design: Qualitative quantitative and mixed methods approaches* (2nd ed.). SAGE Publications.

Crotty, M., Shakespeare, W., & Henry, V. (2020). *The foundations of social researchRCH: Meaning and perspective in the research process*. SAGE Publications. https://doi.org/10.4324/9781003115700

Cypress, B. S. (2017). Rigor or reliability and validity in qualitative research: Perspectives, strategies, reconceptualization, and recommendations. *Dimensions of Critical Care Nursing*, *36*(4), 253–263. https://doi.org/10.1097/DCC.0000000000000253

De Munck, V. C. (2000). Handbook of methods in cultural anthropology. *American Anthropologist*, *102*(1), 183–186. https://doi.org/10.1525/aa.2000.102.1.183

Denzin, N. K. (2017). The Research Act: A Theoretical Introduction to Sociological Methods. In *The Research Act: A Theoretical Introduction to Sociological Methods*. https://doi.org/10.4324/9781315134543

Devotta, K., & Pedersen, C. (2015). Coding qualitative data: Working with a team of coders. *Cultural Anthropology Methods*, *10*(2), 31–36 http://sru.crich.ca

Díaz, J., Pérez, J., Gallardo, C., & González-Prieto, Á. (2023). Applying inter-rater reliability and agreement in collaborative grounded theory studies in software engineering. *Journal of Systems and Software*, *195*, Article 111520 https://doi.org/10.1016/j.jss.2022.111520

Feeley, N., & Gottlieb, L. N. (1998). Classification systems for health concerns, nursing strategies, and Client Outcomes: Nursing practice with families who have a child with a chronic illness. *Canadian Journal of Nursing Research*, *30*(1), 45–60.

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

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., & Grimshaw, J. M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology and Health*, *25*(10), 1229–1245. https://doi.org/10.1080/08870440903194015

Gabay, M. (2017). 21st century cures act. *Hospital Pharmacy*, *52*(4), 264–265. https://doi.org/10.1310/hpj5204-264.

Geertz, C. (2021). Thick description: Toward an interpretive theory of culture [1973]. In *Readings for a History* of *Anthropological Theory, Sixth Edition*.

González-Prieto, Á., Perez, J., Diaz, J., & López-Fernández, D. (2023). Reliability in software engineering qualitative research through Inter-Coder Agreement. *Journal of Systems and Software*, *202*(1) 1–35. https://doi.org/10.1016/j.jss.2023.111707

Graneheim, U. H., & Lundman, B. (2004). Qualitative content analysis in nursing research: Concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today*, *24*(2), 105–112. https://doi.org/10.1016/j.nedt.2003.10.001

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, *18*(1), 59–82. https://doi.org/10.1177/1525822X05279903

Gwet, K. L. (2010). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed.). Advanced Analytics LLC.

Halpin, S. N. (2023). Inter-rater Reliability in Qualitative Coding: Considerations for its Use. https://qualpage.com/2023/08/31/inter-rater-reliability-in-qualitative-coding-considerations-for-its-use/

Halpin, S. N., & Konomos, M. (2022). An iterative formative evaluation of medical education for multiple myeloma patients receiving autologous stem cell transplant. *Journal of Cancer Education* 37 (3), 779-787. https://link.springer.com/article/10.1007/s13187-020-01882-3

Halpin, S. N., Konomos, M., & Jowers, I. (2021). Interrupted identities: Autologous stem cell transplant in patients with multiple myeloma. *Journal of Patient Experience* 8. https://doi.org/10.1177/237437352199886

Halpin, S. N., Dillard, R. L., & Puentes, W. J. (2017). Socio-emotional adaptation theory: charting the emotional process of Alzheimer's disease. *The Gerontologist* 57 (4), 696-706. https://doi.org/10.1093/geront/gnw046

Hoddy, E. T. (2019). Critical realism in empirical research: Employing techniques from grounded theory methodology. *International Journal of Social Research Methodology*, *22*(1), 111–124. https://doi.org/10.1080/13645579.2018.1503400

Joffe, H., & Yardley, L. (2004). Content and thematic analysis. In D. F. Marks & L. Yardley (Eds.), *Research methods for clinical and health psychology* (pp. 56–68). SAGE Publications.

Johnson, J. L., Adkins, D., & Chauvin, S. (2020). A review of the quality indicators of rigor in qualitative research. In *American Journal of Pharmaceutical Education*, *84*(1), Article 7120. https://doi.org/10.5688/ajpe7120

Johnson, R. B., & Onwuegbuzie, A. J. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, *1*(2) 112–133. https://doi.org/10.1177/1558689806298224

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3) 411–433. https://doi.org/10.1093/hcr/30.3.411

Krippendorff, K. (2022). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications. https://doi.org/10.4135/9781071878781

Kurasaki, K. S. (2000). Field methods intercoder reliability for validating conclusions drawn from open-ended interview data. *Field Methods*, *12*(1) 179–194. http://fmx.sagepub.comhttp://fmx.sagepub.com/cgi/content/abstract/12/3/179http://www.sagepublications.comhttp://fmx.sagepub.com/cgi/alertsEmailAlerts:http://fmx.sagepub.com/

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Leung, L. (2015). Validity, reliability, and generalizability in qualitative research. *Journal of Family Medicine and Primary Care*, *4*(3), 324–327. https://doi.org/10.4103/2249-4863.161306

Lincoln, Y. S., Guba, E. G., & Pilotta, J. J. (1985). Naturalistic inquiry. *International Journal of Intercultural Relations*, *9*(4), 438–439. https://doi.org/10.1016/0147-1767(85)90062-8

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. In *Human Communication Research*, *28*(4), 587–604. https://doi.org/10.1093/hcr/28.4.587

Long, H. A., French, D. P., & Brooks, J. M. (2020). Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. Research Methods in Medicine & Health Sciences, *1*(1), 31–42. https://doi.org/10.1177/2632084320947559

MacQueen, K. M., McLellan-Lemal, E., Bartholow, K., & Milstein, B. (2008). Team-based codebook development: Structure, process, and agreement. *Handbook for team-based qualitative research* 119, 119-135.

Marzi, G., Balzano, M., & Marchiori, D. (2024). K-Alpha calculator: Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. *MethodsX*, *12*(1), 1–10. https://doi.org/10.1016/j.mex.2023.102545

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

McAlister, A. M., Lee, D. M., Ehlert, K. M., Kajfez, R. L., Faber, C. J., & Kennedy, M. S. (2017). Qualitative coding: An approach to assess inter-rater reliability. *ASEE Annual Conference and Exposition, Conference Proceedings*. https://doi.org/10.18260/1-2--28777

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. In *Proceedings of the ACM on Human-Computer Interaction*, *72*(3), 1–23. https://doi.org/10.1145/3359174

Miles, M. B., Huberman, M. A., & Saldaña, J. (1994). Qualitative data analysis: A method sourcebook. SAGE Publications.

Moret, M., Reuzel, R., Van Der Wilt, G. J., & Grin, J. (2007). Validity and reliability of qualitative data analysis: Interobserver agreement in reconstructing interpretative frames. In *Field Methods*, *19*(1), 24–39. https://doi.org/10.1177/1525822X06295630

Morse, J. (2020). The changing face of qualitative inquiry. *International Journal of Qualitative Methods*, *19*, 1–7. https://doi.org/10.1177/1609406920909938

Morse, J. M. (1997). "Perfectly healthy, but dead": The myth of inter-rater reliability. *Qualitative Health Research*, *7*(4), 445–447. https://doi.org/10.1177/104973239700700401

Morse, J. M. (2015). Critical analysis of strategies for determining rigor in qualitative inquiry. *Qualitative Health Research*, *25*(9), 1212–1222. https://doi.org/10.1177/1049732315588501

Muskens, G. J. (1980). *Frames of meaning, are they measurable? A methodological critique of the content analysis of illustrated periodical magazines* [Doctoral dissertation, Radbound University]. https://repository.ubn.ru.nl/handle/2066/147940

Nili, A., Tate, M., Barros, A., & Johnstone, D. (2020). An approach for selecting and using a method of inter-coder reliability in information management research. *International Journal of Information Management*, *54*(1), 1–13. https://doi.org/10.1016/j.ijinfomgt.2020.102154

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*, *89*(9), 1245–1251. https://doi.org/10.1097/ACM.0000000000000388

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*, 1–13. https://doi.org/10.1177/1609406919899220

O'Sullivan, T. A., & Jefferson, C. G. (2020). A review of strategies for enhancing clarity and reader accessibility of qualitative research results. *American Journal of Pharmaceutical Education*, *84*(1), Article 7124. https://doi.org/10.5688/ajpe7124

Parker, E. B., & Holsti, O. R. (1970). Content analysis for the social sciences and humanities. *American Sociological Review*, *35*(2), 356–357. https://doi.org/10.2307/2093233

Prasanth, M. (2021). Publication manual of the American Psychological Association: The official guide to APA style. *Kelpro Bulletin*, *25*(2), 90–92.

Roberts, K., Dowell, A., & Nie, J. B. (2019). Attempting rigour and replicability in thematic analysis of qualitative research data: A case study of codebook development. *BMC Medical Research Methodology*, *19*(1), 1–8. https://doi.org/10.1186/s12874-019-0707-y

Roulston, K., & Halpin, S. N. (2022). *Designing qualitative research using interview data*. The SAGE Handbook of Qualitative Research Design. SAGE publications.

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

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

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321–325. https://doi.org/10.1086/266577

Sim, J., Saunders, B., Waterfield, J., & Kingstone, T. (2018). Can sample size in qualitative research be determined a priori? In *International Journal of Social Research Methodology*, *21*(5), 619–634. https://doi.org/10.1080/13645579.2018.1454643

Sword, H. (2015). *The writer's diet: A guide to fit prose*. The University of Chicago Press. https://doi.org/10.7208/chicago/9780226352039.001.0001

Sword, H. (2018). *Air and light and time and space: how successful academics write*. Harvard University Press.

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*(6). 349–357. https://doi.org/10.1093/intqhc/mzm042

Watt, D. (2015). On becoming a qualitative researcher: The value of reflexivity. *The Qualitative Report*, *12*(2), 82–101. https://doi.org/10.46743/2160-3715/2007.1645

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, *13*(1), 1–7. https://doi.org/10.1186/1471-2288-13-61

Woods, M., Paulus, T., Atkins, D. P., & Macklin, R. (2016). Advancing qualitative research using qualitative data analysis software (QDAS)? Reviewing potential versus practice in published studies using ATLAS.ti and NVivo, 1994–2013. *Social Science Computer Review*, *34*(5), 597–617. https://doi.org/10.1177/0894439315596311

Xie, Q. (2013, November). Agree or disagree? A demonstration of an alternative statistic to Cohen's Kappa for measuring the extent and reliability of agreement between observers. In *Proceedings of the Federal Committee on Statistical Methodology Research Conference* (Vol. 4, pp. 294–306). https://nces.ed.gov/FCSM/pdf/J4_Xie_2013FCSM.pdf

Yardley, L. (2000). Dilemmas in qualitative health research. *Psychology and Health*, *15*(2), 215–228. https://doi.org/10.1080/08870440008400302

Zade, H., Drouhard, M., Chinh, B., Gan, L., & Aragon, C. (2018). Conceptualizing disagreement in qualitative coding. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*. https://doi.org/10.1145/3173574.3173733

Zamawe, F. C. (2015). The implication of using NVivo software in qualitative data analysis: Evidence-based reflections. *Malawi Medical Journal*, *27*(1), 13–15. https://doi.org/10.4314/mmj.v27i1.4

Notes on Contributor

Dr. Sean N. Halpin is a Qualitative Analyst with RTI-International, on the Genomics, Ethics, and Translational Research team. Dr. Halpin has over a decade of experience leading socio-behavioral studies across a wide range of chronic and infectious disease areas and has published numerous journal articles to do with patient care. His responsibilities at RTI include preparing research proposals, developing and executing research protocols, overseeing data collection and analysis, interpreting the research results and supporting sponsors' strategic goals, managing operational and financial aspects of research studies, and disseminating results. Dr. Halpin has a Ph.D. in qualitative research and evaluation methodologies from the University of Georgia and an MA in developmental psychology from Teachers College, Columbia University.

ORCID

Sean N. Halpin, https://orcid.org/0000-0001-5624-6083