Article

Sociological Methods & Research
2024, Vol. 53(4) 1944–1975

DOI: 10.1177/00491241231156971
journals.sagepub.com/home/smr


Inter-Rater Reliability Methods in Qualitative Case Study Research

Rosanna Cole[1]

Abstract

The use of inter-rater reliability (IRR) methods may provide an opportunity to improve the transparency and consistency of qualitative case study data analysis in terms of the rigor of how codes and constructs have been developed from the raw data. Few articles on qualitative research methods in the literature conduct IRR assessments or neglect to report them, despite some disclosure of multiple researcher teams and coding reconciliation in the work. The article argues that the in-depth discussion and reconciliation initiated by IRR may enhance the findings and theory that emerges from qualitative case study data analysis, where the main data source is often interview transcripts or field notes. To achieve this, the article provides a missing link in the literature between data gathering and analysis by expanding an existing process model from five to six stages. The article also identifies seven factors that researchers can consider to determine the suitability of IRR to their work and it offers an IRR checklist, thereby providing a contribution to the broader literature on qualitative research methods.

Keywords

inter-rater reliability (IRR), qualitative data, coding, case study, data analysis.

[1] Surrey Business School, University of Surrey, Guildford, UK

Corresponding Author:
Rosanna Cole, Surrey Business School, University of Surrey, Guildford, UK.
Email: r.cole@surrey.ac.uk

Introduction

It is critically important in scholarly work that researchers can demonstrate how they translate their data into findings (or theory) and that the reader has confidence that the findings represent the studied phenomena. This can be particularly challenging in qualitative work, where the reader is often asked to make a leap of faith. This article considers how researchers can use inter-rater reliability (IRR) methods to assess and demonstrate consistency in their analysis of qualitative case study data which could in turn lead to better methodological transparency. Qualitative operations management (OM) research is used as the context for the article, to draw applied examples from a specific literature base. But *beyond* the discipline of OM, the article introduces the IRR process as a supplementary tool in the arsenal of qualitative researchers and explains when using the IRR process may improve the reliability of the analysis and the broader credibility of research results.

Central to all research is the goal of finding plausible and credible outcome explanations using the concepts of reliability and validity to attain rigor (Yin 1994) as "without rigor, research is worthless, becomes fiction, and loses its utility" (Morse et al. 2002:14). The validity of theory or findings *(assuring that what is measured accurately reflects what is intended)* and reliability *(assuring that the data are good readings of the operational measures)* emerging from case research are crucial (McCutcheon and Meredith 1993:246) to provide confidence in the data collected and trust in the application and use of results for managerial decision making (Riege 2003). Although IRR is associated with the consistency of the codes of the data and may not ensure validity, "when it is not established, the data and interpretations of the data can never be considered valid" (Lombard, Snyder-Duch, and Bracken 2002:589). Thus, if the coding is not reliable, the analysis cannot be trusted (Singletary 1993).

Using an IRR process may provide the opportunity to contribute to the improvement of transparency and reliability of data analysis as it is concerned with the procedure of coding the data within a research team, quantitative measures of consistency between coders, and how coding differences are reconciled to demonstrate agreement on the interpretation of the data being coded via consensus building. Having two or more researchers independently analyze the same qualitative data set and then compare their findings, can serve to provide an important check on selective perception and blind interpretive bias. Reliability, as the extent to which a study's operations can be repeated with the same results, is often done in two ways, using a case study protocol and using a case study database (e.g., Stuart et al. 2002). Performing IRR can also allow another researcher to repeat the analytical

coding procedures, beginning with the raw data. Yin (2014) acknowledged the poor documentation of case study research procedures, making external reviewers more suspicious of the reliability of the method. IRR supports the notion of Yin's reliability checks whereby an audit process performs "a reliability check that must be able to produce the same results if the same procedures are followed" (Yin 2014:49), contributing to a quality research design.

IRR may also serve to guard against unethical, opportunistic behavior. Hall and Martin (2019) provided a taxonomy of research misconduct by business school researchers, including data fabrication and falsification, selective reporting, and the omission of data. IRR can provide a check and balance that helps to avoid such misconduct. IRR could contribute to removing any questions of doubt concerning how an article's data has been analyzed. Moreover, it can provide a procedure for systematically challenging and validating the interpretation of the findings from qualitative data. In suitable situations, this can lead to more robust results and may enhance the theory derived from data (e.g., Saldaña 2013; Gwet 2014).

There can be potential for IRR to be used in qualitative research between, and overlapping with, data gathering and analysis. To provide a focus for this article, the qualitative case study method is used (e.g., Eisenhardt 1989; Yin 2017), i.e., the most widely applied qualitative method in the literature where the main data source is often interview transcripts or field notes supported by secondary documents. The use of IRR has been a somewhat controversial subject on the grounds that it may reduce rich, qualitative insight to a series of statistical measures and that it may not allow for differences in interpretation, which can be important to developing more nuanced theory. This article outlines when IRR methods could add the most value and provide a checklist for researchers to use to plot their approach. Even literature on IRR (e.g., Hallgren 2012; Gwet 2014), does not provide detailed guidance on *when* to apply the approach.

Operations Management (OM) as a Context

OM was chosen as a context for this article for two reasons. First, in order to draw on and compare specific samples of literature of IRR, some parameters need to be set. Second, for this article, two seminal papers are used to develop IRR protocols, both from the OM field (Stuart et al., 2002; Barratt, Choi, and Li 2011). For example, this article extends the five-stage OM research process model in Stuart et al. (2002) by inserting a data preparation (and preliminary analysis) stage between data collection and analysis. Further, a detailed 13-step checklist on applying and writing up the use of IRR methods is

developed from the work of Barratt et al. (2011) for qualitative case study research in OM. Their sample of 204 papers is also used to demonstrate a lack of papers employing IRR in OM despite having the data sets to be able to do so. Additionally, in order to seek recent exemplar papers showing a lack of standardization in the application of IRR methods, the OM field was again used.

The remainder of the article is organized as follows. The literature review provides a brief appraisal of papers on the qualitative case study method—the approach focused upon in this article. The Procedure for Assessing IRR section then more formally outlines the approach for incorporating IRR within qualitative data analysis before a checklist to aid researchers in introducing IRR into their own work is provided. The Discussion explains the suitability of IRR for different types of qualitative case research. Finally, conclusions are presented.

Literature Review

The case study method is a key qualitative approach for building theory from the field in OM (e.g., Meredith 1998; Voss, Tsikriktsis, and Frohlich 2002; Barratt et al. 2011). It is especially useful in an emergent sense for new or poorly understood problems as it allows for an in-depth, first-hand investigation of a phenomenon in its natural environment (Edmondson and McManus 2007; Boyer and Swink 2008). But for such qualitative case study research to contribute positively to the development of OM, the method and the theory it develops must stand up to tests of interpretation reliability. Otherwise, the so-called *renaissance* of case study research (Ketokivi and Choi 2014) threatens to be undermined. Thus, in demonstrating how researchers can use measures of IRR in the qualitative case study method, the discussion and reconciliation initiated by IRR may enhance the findings and theory that emerge from qualitative data analysis through interactive and iterative scrutiny of the data.

Despite the potential of the qualitative case study method, there have been many criticisms of the approach and its application in the OM literature. Meredith (1998, citing Aldag and Stearns 1988) noted criticisms of qualitative research methods, such as case studies, including a tendency toward construct error and poor validation. Seuring (2008) called for the case study research process to be more comprehensively documented and Barratt et al. (2011) called for the rigor and consistency of qualitative case work to be elevated—toward greater standardization of the method. Further, both Ketokivi and Choi (2014) and Sodhi and Tang (2014) highlighted the importance of

greater transparency in case research. Finally, in the wider management literature, Pratt (2009) highlighted the importance of showing how authors travel from the data to the findings and Morse et al. (2002) stressed the importance of putting the onus for establishing the reliability (and validity) of qualitative research findings on the authors/researchers rather than the readers/reviewers. We are still some way off a qualitative research culture akin to quantitative methods according to Kuehn and Rohlfing (2022).

Many papers on the case study method have been presented in the OM literature to encourage and guide its application (e.g., McCutcheon and Meredith 1993; Meredith 1998; Stuart et al. 2002; Voss et al. 2002; Barratt et al. 2011). In general, this body of work contributes greatly to our understanding of how to select qualitative cases and collect data, especially via interviews; but pays limited attention to coding and how to use multiple researchers when coding data to improve the reliability of subsequent data analysis. Stuart et al. (2002) presented a five-stage process for qualitative case research as a linear process between data gathering and data analysis. In this article, the model is developed by introducing a new stage between these activities to show the iterative and granular analysis that often actually occurs via within and cross-case analysis, occurring simultaneously and incrementally with data collection (Glaser and Strauss 1967; Vila-Henninger et al. 2022). The literature has acknowledged the value of using coding (Voss et al. 2002, Campbell et al. 2013; Deterding and Waters 2021) and using multiple researchers with qualitative data (e.g., Pratt 2009; Barratt et al. 2011); but there is an opportunity to develop suggestions for how coding teams could operationalize this in practice.

Although there are OM articles on specific aspects of more quantitative empirical approaches, the same does not apply to more qualitative empirical approaches. Pratt (2008) suggests that qualitative data can be more difficult to publish because there are less protocols (which is also part of its attraction). There is an opportunity for qualitative empirical research to demonstrate the same level of rigor and transparency in coding analysis as exhibited by other empirical research methods; and for qualitative case research in particular to demonstrate the same level of rigor and transparency in data coding development and application as it does in case selection and data collection.

It can be concluded that OM scholars have successfully promoted the use of the case method within the discipline over many years; but that the specific issue of IRR has been largely neglected, even when multiple researcher teams are encouraged and there may have been an opportunity to use it. First, there is literature specifically *on* the qualitative case study method, where authors advocate the approach and outline how to conduct rigorous case research.

For example, McCutcheon and Meredith (1993) outlined various aspects of the approach, including case selection and data collection; but the authors put less emphasis on data analysis or preparing the data for analysis via coding. McCutcheon and Meredith (1993) did however highlight the importance of conducting logic tests on the findings of analysis to evaluate the validity and reliability emerging from case work. They suggest that an indication of the researcher's thoroughness bolsters confidence in the findings or indicates shortcomings that may prompt questions about the resulting theory.

Meanwhile, Meredith (1998) discussed the rigor of qualitative case studies in terms of controlled observations, controlled deduction, replicability, and generalizability; but, in this instance, coding or the use of multiple researchers to verify researcher interpretations or inferences was not discussed. The author highlighted the importance of triangulation via multiple data sources but did not refer to triangulation between different coders, which can also add reliability (Patton 1999). Stuart et al. (2002) presented a five-stage case-based research and dissemination process, where Stage 3 is on data gathering and Stage 4 on data analysis. These two stages however need to be bridged by iterative coding and organizing the gathered data for further analysis. The authors highlighted the importance of extracting significant patterns, simplifying descriptive information, and thinking laterally when analyzing case data. They state that how findings are validated is a weakness of case research and called for a more rigorous case research approach. For example, beyond data collection, the key is the interpretation of qualitative information as researchers "make sense from chaos by analyzing and interpreting what individuals are trying to say" (Stuart et al. 2002:427). Further, they briefly referred to developing a database that another researcher could access to repeat the analysis from raw data onwards to improve reliability; but they stopped short of referring explicitly to coding or IRR, which may provide a solid foundation for any creative processes that might follow using discrepancies flagged from coding differences and the resulting reconciliation and consensus building—but not necessarily for the primary goal of agreeing on every code which may erase the value of qualitative work. Meanwhile, Voss et al. (2002) provided an overview of the case study method for OM researchers (see also Voss et al. 2016). Voss et al. (2002) included a discussion on coding to reduce data into categories and on data analysis, but they did not cover IRR. The authors referred to open, axial, and selective coding and within and cross-case analysis.

Second, there is work that *reviews* the use of the qualitative case study method in the extant literature. For example, Seuring (2008) examined the case study application to supply chain management research using Stuart

et al.'s (2002) five-stage model. Seuring (2008) identified a lack of coverage and transparency around aspects of the qualitative case method, including data analysis. That is, how the researcher formulated the overarching themes from the initial participant data. Barratt et al. (2011) then presented an analysis of 204 OM case study papers against 19 evaluation criteria, including justification for qualitative case research, unit of analysis, sampling strategy, number of cases, etc. Only limited attention was given to the data analysis procedures applied with emphasis on within and cross-case analysis. It was beyond the scope of the article to evaluate coding, the number of researchers involved in coding, or the use of IRR methods, but having reviewed the sample of 204 papers, 28% reported that they embarked on some form of collaborative preparation and analysis of the data, i.e., involving multiple researchers, but only 13% of the total papers referred to some detailed aspect of the IRR process (e.g., reconciliation), with only 1% of all papers referring to an IRR statistic (i.e., Vereecke and van Dierdonck 2002; Wu and Choi 2005). This is despite 81% of papers in the sample having some form of textual data that potentially could have been analyzed following an IRR procedure. In doing so, readers of the manuscripts may have experienced stronger confirmation that the analysis was developed from consistent and agreed-upon interpretations of the data. Further, employing an IRR process *may* have further enhanced the findings and theory emerging from this body of work. Although Barratt et al. (2011) mentioned that using multiple researchers can improve confidence in research findings, they did not go into depth on how to operationalize this in practice, e.g., via IRR measures or processes. They briefly referred to an inter-coder agreement in their own analysis of case study papers but did not use this as a criterion to evaluate prior work.

Procedure for Assessing IRR

Interpretive rigor requires the researcher to clearly demonstrate how interpretations of the data have been realized and to illustrate findings with quotations from, or access to, the raw data (Fereday and Muir-Cochrane 2006). Using IRR, a critical indicator is provided that reveals the extent to which independent coders see an artifact and reach the same conclusion (Lombard et al. 2002; Belur et al. 2021). Typically, a set of codes is developed, e.g., into a codebook, often by an individual moving back and forth between the data and a working set of codes, merging, renaming, redefining, or removing codes along the way. The codes may be heavily informed by prior literature and thereby largely deductively derived; informed by the data and literature in

a cyclical approach and thereby largely abductively developed; heavily informed by the data and thereby largely inductively discovered; or the approach may be at some intermediate position between these distinct points.

A sample of the data are then coded by multiple researchers independently before the consistency of their coding is assessed. Have they interpreted the meaning of data in different ways, and why? Have parts of the data been overlooked by one researcher? Does some of the data have more than one meaning? And so on. It is a quantitative indicator of coding reliability—provided as part of a larger process that informs the reconciliation discussion—that should not detract from the qualitative richness and character of the underlying data. Although it is important that research is valid and reliable, Schmenner et al. (2009) cautioned that method should not stifle creativity or innovation, as it is this that can lead to new ideas and theories. The IRR method should be employed to raise areas of dissensus so underlying reasons for this can be investigated, if suitable to do so. Indeed, among other criticisms, there is concern that IRR may stifle creativity and theory building from qualitative data. Yet coding is a key element of qualitative research—"any researcher who wishes to become proficient at doing qualitative analysis must learn to code well and easily. The excellence of the research rests in large part on the excellence of the coding" (Strauss 1987:27). Thus, it is important that coding is proficient and consistent, especially when data are open to interpretation. Coding is partly analytical in itself and forms the foundations for later within and cross-case analysis. Therefore, if coding is not performed appropriately, any subsequent analysis is also undermined. Tired, unskilled, inattentive, and prejudiced coders lead to unreliable results; thus, IRR provides a measure of coding consistency and reliability that seeks to avoid biases and identify mistakes. The IRR process, which forces differences in interpretation to be confronted and encourages consensus building, has the potential to enhance the theory that emerges from data analysis. Before going into greater depth on IRR, Table 1 summarizes how IRR can be used to enhance qualitative research in OM and provides counter-arguments for each one in turn to ensure appropriate application of the method.

IRR Statistics and Evaluation

Hallgren (2012) provided an overview of IRR for observational data, including how to select an appropriate statistic and interpret its value for the field of psychology. The author rejected the use of percentages of agreement as an adequate measure of IRR (citing Cohen 1960; Krippendorff 1980). This

Table 1. What Inter-Rater Reliability (IRR) can do for Qualitative Research.

IRR is …	IRR is not …
About determining the current degree of consensus between coders as a basis for discussion that potentially enhances the findings or theory emerging from the data.	About achieving full consensus, or a score of 1.0, between coders.
About using quantitative measures of reliability to enable the richness of insight that is a key strength of qualitative approaches.	About replacing the richness of insight that is the strength of qualitative approaches with quantitative measures.
About providing transparency and rigor to the creative process of theory building from qualitative data.	About stifling the creative process of theory building from qualitative data.
At times an iterative process of moving back and forth between the data and codes enabled by the IRR procedure.	Necessarily a one-directional confirmation of statistically significant results from data.
About verifying one particular interpretation of the data.	About discovering one single truth in the data.
A mechanical calculation for determining the level of coding agreement between coders that is then supplemented by discussion and reconciliation.	About turning the analysis and write up of qualitative data into a mechanical process.
A measure of agreement that can be calculated manually or automatically using qualitative data analysis software.	Reliant on the use of qualitative data analysis software.

measure does not correct for chance agreements, meaning it overestimates the level of agreement—after all, a stopped clock is right twice a day—and provides no information about statistical power. Hallgren (2012) suggested the most suitable statistic will depend on the metric in which a variable is coded, i.e., whether it is nominal, ordinal, interval, or ratio. In the context of interview transcripts and field notes, nominal data are typical whereby labels or codes are assigned to data extracts or quotations to place them into categories.

Cohen's kappa coefficient (Cohen 1960), which followed on from Scott's *pi* statistic (Scott 1955), is commonly used for nominal (or categorical) variables. It offers a measurement of agreement that could not be expected by chance alone (Cohen 1960) and was used, for example, in the studies by Cole and Aitken (2019, 2020). Kappa statistics measure the observed

agreement level between coders and provide a standardized IRR index that is generalizable across studies. Kappa is scaled between −1.0 and +1.0, where a negative value indicates poorer than chance agreement, zero indicates exactly chance agreement, and a positive value indicates better than chance agreement (Fleiss and Cohen 1973). The upper value, or perfect agreement, of +1.0 indicates that "equally capable coders operating in isolation from each other select the same code for the same unit of text [every time]" (Krippendorff 2004:217). Kappa will rarely take a value below zero unless there is a major disagreement between coders. Hallgren (2012) referred to the many kappa statistic variants presented in the literature, e.g., to handle bias problems or give greater weight to large coding disagreements. For example, Fleiss and Cohen (1973) outlined the weighted kappa statistic and intra-class correlation coefficient (ICC). Further reviews of IRR measures that build on Cohen's kappa are presented, for example, by Barlow, Lai, and Azen (1991), Banerjee et al. (1999), and LeBreton and Senter (2008) for a variety of fields and applications.

Alternative IRR statistics include Krippendorff's alpha, but this is less well-known than Cohen's kappa (and related measures) and is not as widely available in computer programs; and also Perreault and Leigh's (1989) index of reliability procedure for nominal data, which the authors claimed is more suitable than Cohen's kappa for the type of inter-judge data often found in marketing studies. This procedure was used by Wilhelm et al. (2016) and accounts for marginal distribution error, which means that, in certain circumstances, Cohen's kappa may have an upper bound less than +1.0. Similarly, Fleiss' kappa (Fleiss 1971) can be used with nominal data and was applied by Su et al. (2014). This approach can be adopted when there are more than two coders allowing mean IRR statistics to be evaluated for the whole coding team (in addition to evaluating different coding pairs).

Although not without criticism (e.g., Gwet 2002; Krippendorff 2004; Kim et al. 2016), Cohen's kappa (or suitable variants thereof) is the most commonly used approach in the management literature with nominal data and advocated by Hsu and Field (2003). Thus, Cohen's kappa coefficient is the default measurement of choice, as presented in Table 2 together with three key alternatives, e.g., depending on the number of raters and need to account for marginal distribution error. The notation used in the table has been standardized to make the measures more easily comparable. Three of these measures were evident in the OM sample and they form the building blocks of other, more specialized approaches that meet particular research needs, e.g., the weighted approach. The principles of IRR are however

Table 2. Key Alternative Measures and Building Blocks of Inter-Rater Reliability (IRR) for Nominal Qualitative Data Analysis.

IRR statistic	Formulation	Notation	Comments	Key reference	Example of application in OM literature
Scott's (1955) *pi* statistic (π)	$\pi = \frac{p_a - p_c}{1 - p_c}$	p_a= Proportion of agreed judgments between raters. p_c= Probability of chance agreement.	More robust than a simple agreement percentage; used with nominal data; accounts for chance agreement between coders; applies to two coders only.	Scott (1955)	Not identified
Cohen's kappa coefficient (*k*)	$k = \frac{p_a - p_c}{1 - p_c}$	p_a= Proportion of agreed judgments between raters. p_c= Probability of chance agreement.	As above for Scott's (1955) *pi* statistic; formulation appears the same but p_c is calculated differently; unlike in Scott (1955), it is not assumed that coders have the same distribution of responses.	Cohen (1960)	Cole and Aitken (2019)
Fleiss' kappa coefficient (*k*)	$k = \frac{\overline{p}_a - \overline{p}_c}{1 - \overline{p}_c}$	$\overline{p}_a$= Mean agreement between raters. $\overline{p}_c$ = Mean probability of chance agreement.	Builds on Scott's (1955) *pi* statistic rather than Cohen (1960); can be used with three or more coders; applicable to nominal data.	Fleiss (1971)	Su et al. (2014)

(continued)

Table 2. Continued

IRR statistic	Formulation	Notation	Comments	Key reference	Example of application in OM literature
Perreault and Leigh's (1989) index of reliability (I_r)	$I_r = \sqrt{\left(p_a - \frac{1}{c}\right)\left(\frac{c}{c-1}\right)}$	p_a = Proportion of agreed judgements between raters. c = Number of coding categories.	Accounts for marginal distributions that mean Cohen's kappa may have an upper bound <1.0; applies to nominal data.	Perreault and Leigh (1989)	Wilhelm et al. (2016)

OM = operations management.

generic to any chosen statistic, and researchers may choose to report multiple IRR measures.

Cohen's kappa coefficient and some other IRR statistics can be determined manually (see, e.g., Hallgren 2012) or using computer-assisted qualitative data analysis software (CAQDAS). The use of software enhances traceability; it can recognize coding team members and attribute their work to them, providing an audit trail that would be difficult to replicate using manual coding. Coding agreement can first be visually evaluated using coding "stripes", i.e., by comparing the parts of text researchers have highlighted and to which they have assigned codes. This provides an opportunity to observe any major, obvious differences, i.e., where the stripes do not overlap, which can be discussed and the corresponding codes either changed or left unresolved. The level of agreement can then be evaluated more formally by calculating IRR statistics.

Generally, a kappa between +0.61 and +0.80 represents "substantial agreement" and between +0.81 and +1.0 indicates "near perfect agreement" (Landis and Koch 1977). A level of subjective interpretation might be expected and acceptable, thus achieving perfect agreement may not be the objective—it might be argued that divergent views on the data should be embraced and can lead to the emergence of more nuanced theory. Discussions between coders may take place to discuss any divergent views, with the coding team determining what level of agreement should trigger a discussion and what level of agreement should be sought after reconciliation. But, of course, it is important to note that perfect agreement does not guarantee the data has been interpreted correctly—two coders could be equally wrong—it only guarantees the data has been interpreted consistently. Putting a figure on the level of agreement prompts dialogue, which is the real key to further enhancing the findings from qualitative data.

IRR Reconciliation and Reporting

Reconciliation is partly analytical and formative; it can improve the findings or theory emerging from data through consensus building. Campbell et al. (2013) stressed the importance of this stage and of explaining in detail how discrepancies were identified and dealt with. IRR serves to flag where reconciliation may be needed and consensus building could benefit the data analysis. Why did researchers code something differently? Does the data have another, alternative meaning that informs theory development? And so on. It may, for example, be that there are different understandings of how the codes have been defined or how they have been interpreted, there may be

differences in terms of the length of text associated with a code by two researchers, or there may be simple human attention error or differing levels of coding experience, causing discrepancies. Coder experience may include coders at different levels of maturity, degree of coding skill, and/or understanding of the topic. Most importantly, the reconciliation process for addressing differences must be robust and democratic, particularly to avoid the issue of power dynamics in a research team—where senior or more dominant members may influence the reconciliation process and possibly even skew it. For additional validity checks, where possible, findings can be shared with the informants, which may be followed by a further iteration.

In reconciling differences, it may be necessary to go back and adapt the codebook. It is natural for codes to evolve, change, be renamed, or for new codes to be created; and it is important to show how they were merged, absorbed using secondary or tertiary coding categories, and removed; and to present the final set of codes. There are many creative ways of illustrating the coding and IRR journey. In addition to results and reconciliation tables that summarize the presence (or not) of codes (e.g., across cases) and IRR statistics (e.g., overall, by code, and by data source), a schematic can be useful. Saldaña (2013), for example, offered a template to demonstrate the funneling/streamlining of codes in qualitative inquiry.

After reconciliation, and once a satisfactory agreement has been reached, coding can continue either by a single researcher or by multiple researchers in a team-based approach. All of the data could be coded by multiple researchers, but this adds to the workload and may be unnecessary once an agreement has been established for a sizeable sample. Where there is a large amount of data, there are obvious advantages to using multiple researchers to divide up and code the remaining data.

The Overall IRR Process

The above discussion has briefly outlined the process of determining and evaluating IRR in qualitative data analysis, while further guidance can be found in Armstrong et al. (1997), Kurasaki (2000) and Campbell et al. (2013). To summarize, Figure 1 positions the IRR process within Stuart et al.'s (2002) five-stage process model, designed for OM research. The figure extends the model to six stages, placing data preparation (code development and code application) between data gathering and detailed data analysis while also acknowledging that the stages will overlap. Moreover, it breaks the process down into the iterative cycles of code development and code application; and it identifies three distinct approaches for each, i.e.,

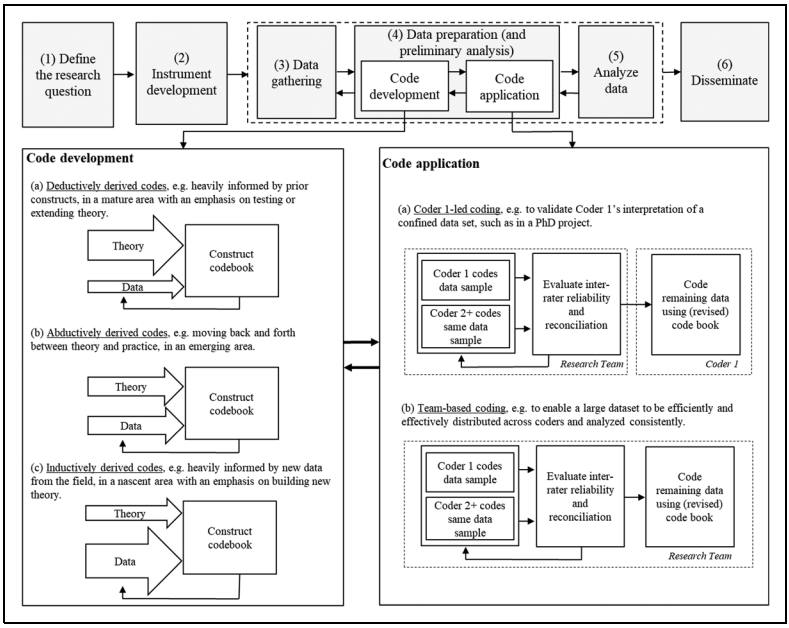

Figure 1. Extending the research process model from Stuart et al. (2002) from five to six stages.

inductively, abductively, and deductively derived codes; and coder 1-led coding and team-based coding. In addition, Figure 2 illustrates in more detail how IRR may be used to evaluate the level of agreement between two or more coders, to reconcile differences, and report on the approach. The following section provides an IRR checklist to support IRR adoption by fellow researchers.

Checklist Development

A 13-step checklist for researchers to consider when embarking on coding qualitative data and writing up their study has been developed. The number of codes generated, the extent of multiple coding and the number of reconciliation rounds required, etc., will inevitably vary from one project to another. Rarely a linear step-by-step procedure is undertaken, even in deductive studies where some preliminary constructs based on the extant literature that describe the phenomenon of interest (*a priori* codes) are used to seek

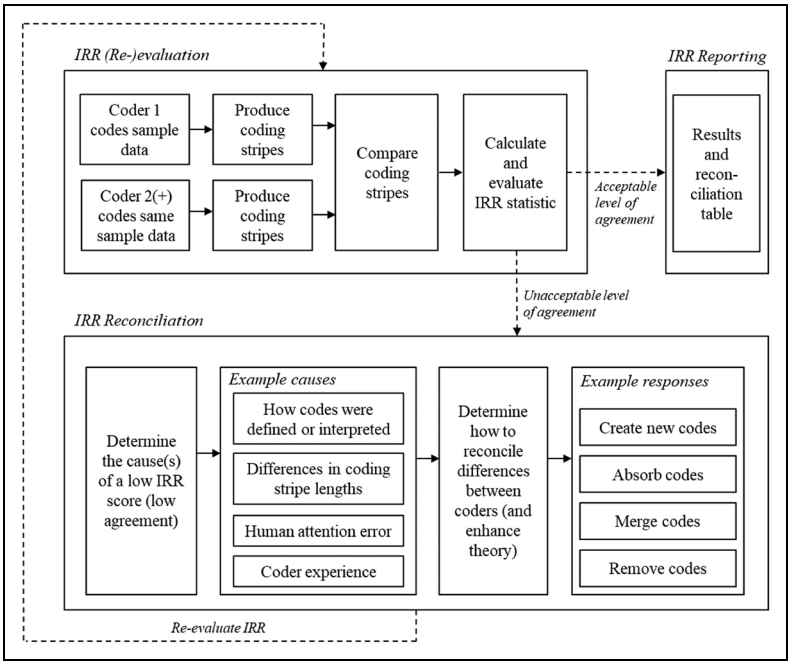

Figure 2. Example of inter-rater reliability (IRR) evaluation, reconciliation, and reporting process.

evidence that addresses these constructs. More often, the research analysis is an iterative and reflexive process and the analysis is guided, not confirmed, by the preliminary codes (Boyatzis 1998). It will also depend on the deductive, inductive, or abductive nature of the research which would determine how the code book is developed (*a priori,* emergent or incremental), whether members of teams should produce separate analyses and then resolve discrepancies or whether joint meetings should generate a single, definitive coded set of materials is a decision for the researcher to make. Although IRR scores lend themselves to deductive studies using *a priori* or pre-defined codes, the IRR process of reconciling for consensus building using IRR score indicators is also important.

After coding has taken place on CAQDAS, using one (or a variation of one) of the types of coding detailed in Figure 1, colored coding stripes can be viewed to allow obviously mismatched codes to be identified visually. Significant mismatches should be discussed to understand why particular

coding options have been taken. Then, either a change to the code or a new agreement on the narrative can be established. For discrepancies, an additional coder is an option, but often these can be resolved through dialogue. At this point, discretion is often used regarding the proportion of matching coding stripes, a potential weakness that is resolved once IRR is calculated, as IRR statistics can direct the coders to the main areas of discrepancy.

IRR scores can be calculated overall and for each construct and data source. During reconciliation, which should be robust and democratic, one of four outcomes usually occurs (see Figure 2) per disagreement: (1) new codes are created; (2) codes are absorbed as a child (second order) into a parent (first order) code; (3) codes are merged to make a new, consolidated code; and (4) codes are removed altogether, where they are deemed no longer necessary. The process improves understanding of the data and takes place because an IRR protocol is being followed.

The research teams can decide how much more data to double code, i.e., at coding saturation, or the point when multiple coding simply becomes a duplication of effort, had been reached. The aim is not to duplicate effort but to determine if coding is consistent and that the approach has multiple verification, even if the last of the data are coded by one person only. IRR statistics can be presented in a table. It may be useful to show codes developing within one case analysis but where codes change across cases, it can be more important to show their development as further cases are analyzed. It is argued here not to be necessary for IRR calculations to continuously improve toward +1.0 as coding is complex. Rather, a context-specific, judgement-informed suitable level of agreement must be reached.

A checklist for conducting IRR-informed analysis is presented in Table 3. Barratt et al. (2011) provided a broad framework toward creating an approach for conducting qualitative case studies for deductive purposes. They support the need to justify the use of cases, state the unit of analysis, explain the sampling procedure, show the number of cases and triangulation of data, and explain data analysis in a logical way. Their paper reinforced the importance of demonstrating transparency in qualitative case research and is utilized by OM scholars. The IRR checklist complements Barratt et al. (2011) by providing 13 steps and IRR considerations relating to the overall process, IRR measures, reconciliation procedure, and subsequent presentation of data. Further, it poses questions a researcher could address when undertaking effective coding and writing up of their work.

It has been highlighted above that IRR is not reported even when declared in many research articles and this may also be because there are challenges in reporting IRR-related information in publications, especially for qualitative

Table 3. Inter-Rater Reliability (IRR) Checklist—13 Steps, Questions for Researchers, and Publication Considerations.

IRR Phase	Step		Questions for the Researchers	Publication Considerations
(Overall Coding and) IRR process	1	Justification of approach	Before embarking on this approach, have you considered other methods of data analysis?	*Have you justified the IRR coding approach?*
	2	Initial coding	Have you identified if you are open coding or using *a-priori* codes, or a combination?	*Have you clearly shown where any a-priori codes come from?*
	3	Production of codebook	Has the first coder produced and disseminated a book of codes and descriptions?	*Have you provided the codebook (e.g., in an appendix)?*
	4	Training of coding team	Has the coding team been trained in identifying and recording codes (and any related software)?	*Have you explained how the coding team were chosen?*
	5	Proportion of multiple coding	Have you decided when you will stop multiple coding and assess IRR?	*Have you explained your justification for coding saturation being reached?*
IRR measures	6	Analysis of coding comparisons	Have you considered the different methods for comparing IRR (e.g., coding stripes and the various statistics)? Have you considered the length of coding stripes for passages?	*How will you present your coding work?*
	7	IRR statistics reported	Have you decided what types of IRR statistics to present and in what format?	*Have you presented the IRR statistics?*
	8	Acceptable levels of IRR	Has the acceptable level of IRR been determined? Are you showing a movement from weak to strong codes? Are you demonstrating poor codebook development that needs strengthening or complex constructs?	*Has an explanation of the IRR allowances been given, with justification?*
Reconciliation of IRR	9	Details of reconciliation	Have you developed a process for reconciliation? Does	*Is the reconciliation process explained?*

(continued)

Table 3. Continued

IRR Phase	Step	Questions for the Researchers	Publication Considerations
differences (and presentation of results)		the coding team understand the process for reconciling differences?	
	10 Details of absorption	Have you decided how you will interpret first and second order code constructs and categories? And, have you decided on the options for 'rejected' codes?	*Have you presented the development of the codes through the process?*
	11 Participant validity	Have you invited your subjects to validate your results?	*How will you present the iterative nature of the validity?*
	12 Triangulation of data sources	Have you considered how the coding protocol might change for different types of data (e.g., interview and document)? Have you considered how you will present the analysis of different types of data sets?	*How will you present different types of data sources coded in the same way?*
	13 Presentation of results	Have you considered different methods of presenting coded data?	*Have you presented the codes and IRR statistics in the most effective way?*

research papers where word counts can be very tight for papers reporting qualitative research. More often than not, methodological details such as how coding developed and changed resulting from IRR, are sacrificed in the interests of reporting findings—thus providing less incentive to researchers to take the time and effort to conduct IRR. However, there is an increasing pressure on researchers to make their data and analysis publicly available and open to scrutiny. For example, the Research Councils UK (RCUK) open access policy asks for publicly funded research to include a statement in any research output concerning how readers can access the underlying research materials, such as data, samples, or models. Journal publishers are also encouraging authors to write a Data Availability Statement (Sage and Wiley) or Data In Brief (Elsevier), revealing the data and how it has been analyzed, or making the underlying data available in a data repository to provide readers with greater understanding of the research.

Discussion: Determining the Suitability of IRR Methods

In this article, the use of IRR methods has been discussed, with a particular focus on the qualitative case study method, which is an important approach for building theory and understanding new or under-explored problems. Indeed, Edmondson and McManus (2007) argued the case study method is well-suited when theory is at a relatively nascent development stage. Such inductive work may rely on developing new, rather than using existing constructs or measures. This opens up the opportunity to make new discoveries and original contributions, but it also naturally adds uncertainty and complexity to coding and data interpretation. Ketokivi and Choi (2014) later unpacked three more specific case study modes with differing degrees of emphasis on the empirical context and general theory, i.e., theory generation, testing, and elaboration. This article has argued that IRR methods are applicable when generating, testing and elaborating theory; however, it may be more crucial to demonstrate reliability when generating new rather than testing established theory as theory generation is likely to be more prone to subjective bias. Meanwhile, the maturity of the general theory base upon which the research is built and the nature of the study, being inductive, abductive or deductive, may affect the stability of the initial codebook, the number of iterations it must go through, and what is considered an achievable IRR level. Thus, a stable codebook and high level of IRR may be achieved quickest when the research builds on an established body of literature and known constructs. But, paradoxically, IRR may be particularly important when developing emergent codes, confirming interpretations of reality through a blind analysis

and consensus building. That said, it is much trickier, would not be suitable for some studies (e.g., some ethnography) and would attract much more criticism from qualitative researchers who believe IRR would stifle the creative process.

Nonetheless, IRR methods will fit with some approaches more than others. For example, Pratt (2009) acknowledged that multiple coders may make little sense in a deeply personal and ethnographic study as other researchers could not develop an adequate contextual understanding to code data in a meaningful way. In contrast, the approach would be well-suited to archival data that has not been directly collected for the purposes of the research and does not rely on a deeply personal understanding of the context. Thus, the data source, and the importance of understanding both the context in which the data are situated and how it was collected become important considerations in determining if coding reliability needs to be established and whether multiple researchers could reasonably be expected to code or interpret data they have not collected themselves.

Thus, there is a challenge to the whole notion of consistency in analyzing data (depending on how one views reality)—in that it is not possible to represent the social world as it means different things to different people. However, Glaser and Strauss (1967) claimed that the virtue of even inductive processes was that they ensured the theory was closely related to the daily realities of what is actually occurring. Many scholars agree that there needs to be some interpretation consistency and suggest using multiple researchers to verify the interpretation of data to do this (e.g., Patton 1999). Yin (2014) highlighted the importance of examining plausible rival explanations of case study data through multiple researchers and Eisenhardt (1989) explained that multiple investigators can foster divergent perspectives and strengthen the grounding of theory. Meanwhile, Miles, Huberman and Saldaña (2014) advocated team coding for definitional clarity and reliability and Saldaña (2013) outlined how collaborative coding enables the power of multiple minds and ways of analyzing or interpreting data. The latter author however cautioned that it is important to find a way of coordinating and harmonizing individual coding efforts—IRR is a means of achieving this. The debate still runs on about whether researchers should be expected to find the same answers or not (Armstrong et al. 1997) but a theoretical resolution of these divergent positions is impossible as their core ontological assumptions are so different.

Although the focus of this article has been on qualitative case study research (and the most likely analysis of interview transcripts or field notes), IRR methods can be used with other qualitative methods and data

Table 4. An Assessment of the Suitability of the Inter-Rater Reliability (IRR) Approach.

Condition		Assessment of IRR Suitability	Justification
Maturity of the field	Mature	Highly suitable	IRR can be used irrespective of the maturity of the field; but in more mature fields of research, where there is greater prior understanding of a phenomenon, researchers may arrive at a stable working set of codes more quickly making coding and reaching agreement arguably more straightforward. Thus, the maturity of the field does not affect if IRR can be used but it may affect, to some degree, how it is used or the number of iterations.
	Nascent	Highly suitable	
Type of theory application	Building	Highly suitable	IRR can be used with a range of theory applications, including theory building and theory testing approaches. The process however may differ according, for example, to the deductive versus inductive nature of the research (e.g., whether *a priori* codes are derived from the literature or if codes emerge from the data—also closely linked to the maturity of the field).
	Testing	Highly suitable	
Availability of data	High	Highly suitable	Some form of data are necessary for IRR to be applied, most typically in text format, e.g., interview transcripts, detailed interview notes, diary entries, newspaper reports, company reports, etc. A greater amount of data arguably leads to greater complexity,
	Low	Unsuitable	

(continued)

Table 4. Continued

Condition		Assessment of IRR Suitability	Justification
			which may mean there is more cause to check the consistency of the interpretation.
Number of types of data	Multiple	Highly suitable	IRR supports the analysis of singular or multiple data sources. In the context of multiple data sources, it enables them to be compared and integrated providing the same codebook can be used across data sources or formats.
	Singular	Highly suitable	
Reliance on the coding team understanding the context	Low	Highly suitable	When there is a low reliance on understanding the context, multiple coders can code and interpret the text irrespective of whether they were involved in the collection of any primary data. In abductive reasoning, understanding the discipline is often unnecessary as by its very nature abductive coding can go out of the predominant theoretical field—that is, the highly iterative nature of the process may take researchers down any theoretical route. In deeply ethnographic studies, it becomes more difficult for multiple coders to be involved in coding and validation as there can be a strong need to understand the idiosyncratic context in which qualitative data has been collected.
	High	Unsuitable *unless coding team are involved in data collection*	
Power structure (composition	Autocratic	Highly suitable	IRR benefits from an autocratic and democratic

(continued)

Table 4. Continued

Condition		Assessment of IRR Suitability	Justification
and organization of research team)	Democratic	Highly suitable	reconciliation process where power dynamics are fair and equal. Unequal power dynamics are likely to prove problematic for IRR which needs the research team views to be considered equal to avoid skewing the data toward the views of the senior member(s). Research teams often boast a principal investigator but this does not necessarily mean a hierarchical composition of the research team in terms of power dynamic. Due to the nature of the iterative process needed, IRR is unlikely to prove successful in an unstructured culture of a chaotic team composition.
	Hierarchical	Unsuitable	
	Chaotic	Unsuitable	
Use of qualitative data analysis software	Yes	Highly suitable	IRR does not rely on using qualitative data analysis software; but the use of such software improves the audit trail and it also makes sharing data, involving multiple (potentially geographically dispersed) coders, and (re-)calculating the IRR statistic quicker and more straightforward. Thus, the use of software is advised but is not a prerequisite for the IRR approach.
	No	Suitable	

forms (Belur et al. 2021). For example, IRR could be used in conjunction with analyzes of news reports, internal company documents (e.g., policies, contracts, etc.), websites, open-ended survey questions, diaries, etc.; or even audio data using CAQDAS. Using multiple methods is supportive of a shift away from methodological tribalism (e.g., Saunders and Bezzina 2015) and means a more complete and reliable understanding of a phenomenon can be established (Boyer and Swink 2008). Of course, a mixed methods approach may be adopted whereby, for example, qualitative and quantitative data are combined (e.g., Boyer and Swink 2008; Choi, Cheng, and Zhao 2016; Huarng et al. 2018) for the purpose that the two (or more) approaches complement one another, such as in qualitative comparative analysis (Lo, Rey-Martí, and Botella-Carrubi 2020). Guercini (2014) promoted "hybridization" between different qualitative methods and between qualitative and quantitative methods. But hybridization has also been used to refer to transforming qualitative data into meaningful quantitative results, i.e., from words to numbers (Srnka and Koeszegi 2007; Lo et al. 2020). This, however, could arguably dilute the value of having rich qualitative data (Pratt 2009). Thus, IRR methods can be used to provide a level of transparency and reliability to qualitative analysis, where the statistics should not replace insight but promote conversations about the data. Using IRR to determine the accuracy or validity of data is amiss and authors should not be inclined to use it as such or be convinced into doing so by reviewers. It should be noted that plenty of other variables can bring the rigor of qualitative case study data into question, such as poor case selection, confusion in the unit of analysis, sloppy sampling logic.

Finally, authors hoping to enhance the generalizability of qualitative research using IRR should take heed. Whereas with deductive quantitative research, the aim is empirical generalizability and empirical replication, in inductive qualitative research the aim is conceptual generalizability and concepts do not achieve their value by replication but by usefulness. The suitability of the use of IRR in qualitive case study research must be considered carefully by the researcher. Table 4 summarizes the suitability of the IRR approach to researchers' work and can be used when justifying a decision either way.

Conclusions

This article has outlined how and when researchers may consider using IRR methods to attempt to improve the transparency and reliability of qualitative data analysis, reducing the risk of misinterpreting the meaning of data and

building consensus of the data analysis within their study. Authors should deliberate what they hope to achieve using IRR and assess the suitability of their study. That said, the consistency of coding measure may lead to an increase in rigor (which *may* in turn lead to an increase in validity) by determining the applicability of the code to the raw information and reducing interpretation bias for projects where the results are used for managerial decision making (Riege 2003). Although reliability and validity refer to different research outcomes and IRR by its very title is a measure of reliability (through the ability of the transparent process to offer ease in the replication of the analysis), Kvale (1989) states that to validate is to investigate, to check, to question, and to theorize. All of these activities (of which IRR is a part) are integral components of qualitative inquiry that enhance rigor, which in turn achieve validity (Morse et al., 2002). To provide a focus, IRR for the case study method has been exercized, i.e., the most commonly applied qualitative research approach in the literature. Although scholars have offered exhaustive guides on how to select cases and collect data, far less attention has been paid to how to prepare and analyze the acquired data.

The five-stage process model in Stuart et al. (2002) has been extended by inserting a data preparation (and preliminary analysis) stage between data collection and analysis. In Figure 1, this is split into two overlapping and interrelated parts, i.e., code development and code application. Although other intermediate and hybrid points will exist, the figure outlines three typical approaches to each part—that is, inductively, abductively, and deductively derived codes; and coder 1-led coding and team-based coding. In doing so, the figure provides a greater level of granularity to the IRR process than can be found in the extant literature. Further, the article has provided a detailed 13-step checklist on applying and writing up the use of IRR methods.

Overall, IRR comes up against key challenges in its usefulness and acceptance in case research, which is often chosen for its capability to stimulate creativity and harvest rich insights. It must be noted that IRR is not a silver bullet to make research "better", in response to criticisms of construct error and poor transparency of how data led to findings. For example, prior to embarking on the IRR process, the research design and data collection will need to be of high quality. IRR enhances some projects where consistency in interpretation, thus reducing interpretation bias, needs to be mitigated. It makes the journey from raw data to analysis more rigorous (through reconciliation discussion which is often implicit in qualitative research papers) and transparent (through explicit communication of the

IRR measure) but does not *necessarily* lead to validity which would accurately reflect the phenomenon. Also, IRR statistics provide proof of the consistency of coding decisions but do not identify the quality of those decisions, or the degree of democracy in the reconciliation process. Finally, design and implementation decisions are often impacted by the researcher's opinions, preferences, and beliefs, that is, after all, the point of epistemology, ontology and axiology.

The article has argued though that there may have been a missed opportunity in the methods literature as the IRR process and statistic could provide a platform for discussion that has the potential to enhance the findings and theory emerging from qualitative data analysis through reconciliation, discussion, and consensus building. Although IRR is not a guarantee of good research, it does not, for example, have to stifle creativity or turn qualitative data analysis into a mechanical process, especially as many qualitative data sets would not suit it. IRR is not about taking the richness out of qualitative research or reducing it down to a set of statistics. Rather, it is about providing an audit trail that gives the reader confidence in the process the researchers have gone through to unpack and reliably make sense of their rich data, allowing it to be heard and enabling the full potential of the method. There may also be a particular opportunity to augment research methods training with IRR considerations to encourage future research leaders to further enhance the rigor of qualitative data analysis.

Data Availability Statement

Data availability is not applicable to this article as no new data were created or analyzed in this study.

Declaration of Conflicting Interests

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

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Rosanna Cole https://orcid.org/0000-0002-7744-1942

References

Aldag, Ramon J. and Timothy M. Steams. 1988. "Issues in Research Methodology." *Journal of Management* 14(2):253-76.

Armstrong, David, Ann Gosling, John Weinman, and Theresa Marteau. 1997. "The Place of Inter-Rater Reliability in Qualitative Research: an Empirical Study." *Sociology* 31(3):597-606.

Banerjee, Mousumi, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. 1999. "Beyond Kappa: a Review of Interrater Agreement Measures." *Canadian Journal of Statistics* 27(1):3-23.

Barlow, William, Mei-Ying Lai, and Stanley P. Azen. 1991. "A Comparison of Methods for Calculating a Stratified Kappa." *Statistics in Medicine* 10(9):1465-72.

Barratt, Mark, Thomas Y. Choi, and Mei Li. 2011. "Qualitative Case Studies in Operations Management: trends, Research Outcomes, and Future Research Implications." *Journal of Operations Management* 29(4):329-42.

Belur, Jyoti, Lisa Tompson, Amy Thornton, and Miranda Simon. 2021. "Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making." *Sociological Methods & Research* 50(2):837-65.

Boyatzis, Richard. E. 1998. *Transforming Qualitative Information: Thematic Analysis and Code Development*. USA: Sage Publications, Inc.

Boyer, Kenneth K. and L. Swink Morgan. 2008. "Empirical Elephants – Why Multiple Methods are Essential to Quality Research in Operations and Supply Chain Management." *Journal of Operations Management* 26:337-48.

Campbell, John L., Charles Quincy, Jordan Osserman, and Ove K. Pedersen. 2013. "Coding in-Depth Semistructured Interviews: problems of Unitization and Intercoder Reliability and Agreement." *Sociological Methods & Research* 42(3):294-320.

Choi, Tsan-Ming, T. C. E. Cheng, and Xiande Zhao. 2016. "Multi-methodological Research in Operations Management." *Production and Operations Management* 25(3):379-89.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20(1):37-46.

Cole, R. and J. Aitken. 2019. "Selecting Suppliers for Sustainable Supply Chain Management: post-Exchange Supplier Development Activities as pre-Selection Requirements." *Production Planning and Control* 30(14):1184-202.

Cole, R. and J. Aitken. 2020. "The Role of Intermediaries in Achieving Sustainable Supply Chain Management." *Journal of Purchasing and Supply Management* 26(2).

Deterding, Nicole M. and Mary C. Waters. 2021. "Flexible Coding of in-Depth Interviews: a Twenty-First-Century Approach." *Sociological Methods & Research* 50(2):708-39.

Edmondson, Amy. C. and E. McManus Stacy. 2007. "Methodological fit in Management Field Research." *Academy of Management Review* 32(4):1246-64.

Eisenhardt, Kathleen. M. 1989. "Building Theories from Case Study Research." *Academy of Management Review* 14(4):532-50.

Fereday, Jennifer and Eimear Muir-Cochrane. 2006. "Demonstrating Rigor Using Thematic Analysis: a Hybrid Approach of Inductive and Deductive Coding and Theme Development." *International Journal of Qualitative Methods* 5(1):80-92.

Fleiss, Joseph. L. 1971. "Measuring Nominal Scale Agreement among Many Raters." *Psychological Bulletin* 76(5):378-82.

Fleiss, Joseph L. and Jacob Cohen. 1973. "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability." *Educational and Psychological Measurement* 33(3):613-9.

Glaser, Barney G. and L. Strauss Anselm. 1967. *The Discovery of Grounded Theory, Strategies for Qualitative Research*. Hawthorne, NY: Aldine de Gruyter.

Guercini, Simone. 2014. "New Qualitative Research Methodologies in Management." *Management Decision* 52(4):662-74.

Gwet, Kilem. 2002. "Kappa Statistic is not Satisfactory for Assessing the Extent of Agreement Between Raters." *Statistical Methods for Inter-Rater Reliability Assessment* 1(6):1-6.

Gwet, Kilem L. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters*. Advanced Analytics, LLC.

Hall, Jeremy. and R. Martin Ben. 2019. "Towards a Taxonomy of Research Misconduct: the Case of Business School Research." *Research Policy* 48(2):414-27.

Hallgren, Kevin. A. 2012. "Computing Inter-Rater Reliability for Observational Data: an Overview and Tutorial." *Tutorials in Quantitative Methods for Psychology* 8(1):23.

Hsu, Louis. M. and Ronald Field. 2003. "Interrater Agreement Measures: comments on Kappan, Cohen's Kappa, Scott's π, and Aickin's α." *Understanding Statistics* 2(3):205-19.

Huarng, Kun-Huang, Andrea Rey-Martí, and María-José Miquel-Romero. 2018. "Quantitative and Qualitative Comparative Analysis in Business." *Journal of Business Research* 89:171-4.

Ketokivi, Mikko. and Thomas Choi. 2014. "Renaissance of Case Research as a Scientific Method." *Journal of Operations Management* 32(5):232-40.

Kim, Sang-Yeon, Scott S. Graham, Seokhoon Ahn, Michele K. Olson, Daniel J. Card, Molly M. Kessler, Danielle M. DeVasto, Laura R. Roberts, and Fallon A. Bubacy. 2016. "Correcting Biased Cohen's Kappa in NVivo." *Communication Methods and Measures* 10(4):217-32.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to its Methodology*. Beverly Hills, CA: Sage Publications.

Krippendorff, Klaus. 2004. "Reliability in Content Analysis." *Human Communication Research* 30(3):411-33.

Kuehn, D. and I. Rohlfing. 2022. "Do Quantitative and Qualitative Research Reflect two Distinct Cultures? An Empirical Analysis of 180 Articles Suggests "no." *Sociological Methods & Research*, *in press.*

Kurasaki, K. S. 2000. "Intercoder Reliability for Validating Conclusions Drawn from Open-Ended Interview Data." *Field Methods* 12(3):179-94.

Kvale, Steinar. 1989. *Issues of Validity in Qualitative Research*. Lund, Sweden: Chartwell Bratt.

Landis, J., G. Richard, Gary, and Koch, 1977. "The Measurement of Observer Agreement for Categorical Data", *Biometrics* 33(1):159-74.

LeBreton, James. M. and L. Senter Jenell. 2008. "Answers to 20 Questions About Interrater Reliability and Interrater Agreement." *Organizational Research Methods* 11(4):815-52.

Lo, Fang-Yi, Andrea Rey-Martí, and Dolores Botella-Carrubi. 2020. "Research Methods in Business: Quantitative and Qualitative Comparative Analysis." *Journal of Business Research*: 221-24.

Lombard, Matthew, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2002. "Content Analysis in Mass Communication: assessment and Reporting of Intercoder Reliability." *Human Communication Research* 28(4):587-604.

McCutcheon, David. M. and Jack. R. Meredith. 1993. "Conducting Case Study Research in Operations Management." *Journal of Operations Management* 11(3):239-56.

Meredith, Jack. 1998. "Building Operations Management Theory Through Case and Field Research." *Journal of Operations Management* 16(4):441-54.

Miles, Matthew B., A. Michael Huberman, and Johnny Saldaña. 2014. *Qualitative Data Analysis*. Thousand Oaks, California, USA: Sage Publications, Third Edition.

Morse, Janice M., Michael Barrett, Maria Mayan, Karin Olson, and Jude Spiers. 2002. "Verification Strategies for Establishing Reliability and Validity in Qualitative Research." *International Journal of Qualitative Methods* 1:13-22.

Patton, Michael Quinn. 1999. "Enhancing the Quality and Credibility of Qualitative Analysis." *Health Services Research* 34(5, Part II):1189-208.

Perreault, D. William Jr and Laurence E. Leigh. 1989. "Reliability of Nominal Data Based on Qualitative Judgments." *Journal of Marketing Research* 26(2):135.

Pratt, Michael. G. 2008. "Fitting Oval Pegs into Round Holes: tensions in Evaluating and Publishing Qualitative Research in top-Tier North American Journals." *Organizational Research Methods* 11(3):481-509.

Pratt, Michael. G. 2009. "From the Editors: for the Lack of a Boilerplate: tips on Writing up (and Reviewing) Qualitative Research." *Academy of Management Journal* 52(5):856-62.

Riege, Andreas. M. 2003. "Validity and Reliability Tests in Case Study Research: a Literature Review with "Hands-on" Applications for Each Research Phase." *Qualitative Market Research: An International Journal* 6(2):75-86.

Saldaña, Johnny. 2013. *The Coding Manual for Qualitative Researchers*. London: Sage Publications, Ltd.

Saunders, Mark. N. and Frank. Bezzina. 2015. "Reflections on Conceptions of Research Methodology among Management Academics." *European Management Journal* 33(5):297-304.

Schmenner, Roger W., Luk Van Wassenhove, Mikko Ketokivi, Jeff Heyl, and Robert F. Lusch. 2009. "Too Much Theory, not Enough Understanding." *Journal of Operations Management* 27(5):339-43.

Scott, William, A. 1955. "Reliability of Content Analysis: the Case of Nominal Scale Coding." *Public Opinion Quarterly* XIX:321-5.

Seuring, Stefan. A. 2008. "Assessing the Rigor of Case Study Research in Supply Chain Management." *Supply Chain Management: An International Journal* 13(2):128-37.

Singletary, Michael. W. 1993. *Mass Communication Research: Contemporary Methods and Applications*. Boston: Addison-Wesley.

Sodhi, ManMohan S. and Christopher S. Tang. 2014. "Guiding the Next Generation of Doctoral Students in Operations Management." *International Journal of Production Economics* 150:28-36.

Srnka, Katharina J. and Sabine T. Koeszegi. 2007. "From Words to Numbers: how to Transform Qualitative Data into Meaningful Quantitative Results." *Schmalenbach Business Review* 59(1):29-57.

Strauss, Anselm. L. 1987. *Qualitative Analysis for Social Scientists*. Cambridge: Cambridge University Press.

Stuart, Ian, David McCutcheon, Robert Handfield, Ron McLachlin, and Danny Samson. 2002. "Effective Case Research in Operations Management: a Process Perspective." *Journal of Operations Management* 20(5):419-33.

Su, Hung-Chung, Kevin Linderman, Roger G. Schroeder, and Andrew H. Van de Ven. 2014. A Comparative Case Study of Sustaining Quality as a Competitive Advantage." *Journal of Operations Management* 32(7):429-45.

Vereecke, Ann. and Roland Van Dierdonck. 2002. "The Strategic Role of the Plant: Testing Ferdows's Model." *International Journal of Operations and Production Management* 22(5):492-514.

Vila-Henninger, Luis, Claire Dupuy, Virginie Van Ingelgom, Mauro Caprioli, Ferdinand Teuber, Damien Pennetreau, Margherita Bussi, and Cal Le Gall. 2022. "Abductive Coding: theory Building and Qualitative (Re) Analysis." *Sociological Methods & Research*. In press.

Voss, Christopher., Johnson Mark, and Godsell Jan 2016. "Case Research in Operations Management." Pp. 165–97 in *Research Methods for Operations Management*, edited by C. Karlsson. London: Routledge.

Voss, Christopher., Nikos Tsikriktsis, and Markham Frohlich. 2002. "Case Research in Operations Management." *International Journal of Operations & Production Management* 22(2):195-219.

Wilhelm, Miriam M., Constantin Blome, Vikram Bhakoo, and Antony Paulraj. 2016. "Sustainability in Multi-Tier Supply Chains: understanding the Double Agency Role of the First-Tier Supplier." *Journal of Operations Management* 41:42-60.

Wu, Zhaohui and Thomas Y. Choi. 2005. "Supplier–Supplier Relationships in the Buyer–Supplier Triad: building Theories from Eight Case Studies." *Journal of Operations Management* 24(1):27-52.

Yin, Robert. K. 1994. "Discovering the Future of the Case Study Method in Evaluation Research." *Evaluation Practice* 15:283-90.

Yin, Robert. K. 2014. *Case Study Research and Applications: Design and Methods*. Thousand Oaks, California, USA: Sage Publications.

Yin, Robert. K. 2017. *Case Study Research and Applications: Design and Methods*. Thousand Oaks, California, USA: Sage Publications.

Author Biography

Rosanna Cole is a senior lecturer in Sustainable Supply Chain Management at Surrey Business School, University of Surrey, Guildford, UK. She uses mostly qualitative research methods and publishes her coding and inter-rater reliability teamwork protocols in detail. She is an associate editor for the European Management Journal where she supports work on all aspects of operations and supply chain management studies using a range of methodological approaches.