International Journal of Information Management 54 (2020) 102154

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

An approach for selecting and using a method of inter-coder reliability in information management research

Alireza Nili[a,*], Mary Tate[b], Alistair Barros[a], David Johnstone[b]

[a] Queensland University of Technology, Australia
[b] Victoria University of Wellington, New Zealand

ARTICLE INFO

Keywords:
Inter-coder reliability
Content analysis
Information
Management

ABSTRACT

Qualitative researchers in information management research often need to evaluate inter-coder reliability (ICR) to test the trustworthiness of their content analysis. A suitable method of evaluating ICR enables researchers to rigorously assess the degree of agreement among two or more independent qualitative coders. This allows researchers to identify mistakes in the content analysis before the codes are used in developing and testing a theory or a measurement model and avoid any associated time, effort and financial cost. Different methods have been proposed, but little guidance is available on which approach to evaluating ICR should be used. In this paper, we review and compare leading ICR methods that are suitable for qualitative information management research. We propose an approach for selecting and using an ICR method, supported by an illustrative example. The five steps in our proposed approach include: selecting an ICR method based on its characteristics and requirements of a project; developing a coding scheme; selecting and training independent coders; calculating the ICR coefficient and resolving discrepancies; and reporting the process of evaluating ICR and its results.

1. Introduction

Leading information management journals have recognized that information management research needs to focus on understanding and changing human behavior, and the way people use information to engage with knowledge-focused activities.[1] This means that research devoted to understanding and explaining the rich experience of information users is receiving increasing focus.[2] Many information management studies employ qualitative methods such as semi-structured interviews and focus groups with information users, and content analysis of these data is an important part of these studies (Boudreau, Gefen, & Straub, 2001; Davies, 2012). In content analysis, data in the transcripts (audio and/or video records of participants' responses converted into text) are typically coded by trained qualitative coders, and these codes can be trusted only after ensuring their reliability (Davies, 2012; Venkatesh, Brown, & Bala, 2013). To ensure the reliability of the coding, a method of evaluating Inter-Coder Reliability (ICR) needs to be employed.

However, there is currently no agreement, and little guidance for researchers, in selecting and applying ICR methods. Although evaluating ICR is a significant component of content analysis in studies of information management, selecting and using the right method is still a significant challenge for researchers (Davies, 2012; Olson, McAllister, Grinnell, Walters, & Appunn, 2016; Venkatesh et al., 2013), and "researchers who conduct qualitative research have long faced the challenge of providing appropriate reliability" (Park & Park, 2015, p. 180). Also, information on these methods and the process of using them is diffuse, and frequently obtained from publications in other disciplines, such as health, education and media studies. Nevertheless, reviewers and editors expect an ICR check for qualitative studies. There is a

* Corresponding author.
E-mail addresses: a.nili@qut.edu.au (A. Nili), mary.tate@vuw.ac.nz (M. Tate), alistarir.barros@qut.edu.au (A. Barros), david.johnstone@vuw.ac.nz (D. Johnstone).

[1] In recent information management research "there is greater focus on managing activities that make changes in patterns of behavior of customers, people, and organizations, and information that leads to changes in the way people use information to engage in knowledge-focused activities" (International Journal of Information Management, 2019). See https://www.journals.elsevier.com/international-journal-of-information-management

[2] Examples of such research, in emergent contexts, include: perceptions and attitudes towards Blockchain technology (Hughes, et al., 2019), Artificial intelligence, big data analytics and challenges for decision making (Dwivedi et al., 2019), research on Smart cities (James et al., 1984; Ismagilova, Hughes, Dwivedi, & Raman, 2019), the role of hedonic motivation in information systems use (Tan et al., 2016; Tamilmani, Rana, Prakasam, & Dwivedi, 2019), and the role of social media for warning in disasters (Zhang, Fan, Yao, Hu, & Mostafavi, 2019).

https://doi.org/10.1016/j.ijinfomgt.2020.102154
Received 3 July 2019; Received in revised form 9 April 2020; Accepted 20 May 2020

disconnect between the espoused practice (which is to employ rigorous evaluation of ICR) in information management research and the actual practice which is frequently more scattered. We answer the following research questions:

RQ1: What are the criteria that researchers need to consider when selecting the most suitable method for evaluating ICR that meets the specific characteristics of their research study?; and

RQ2: How should an ICR method be applied?

Evaluating ICR enables researchers to assess the degree of agreement between two or more independent coders on the data chunks (e.g. relevant participants' comments, responses or opinions) in the transcript. The more coders agree on the codes, the more confident researchers can be that the codes used by one coder are exchangeable with codes provided by another, and therefore, the findings are reproducible and trustworthy (Davies, 2012; Olson et al., 2016). By employing a suitable ICR method, researchers can identify and correct mistakes in the content analysis before the codes are used in developing a theory, theoretical framework, or a measurement model, and so avoid: errors, costs in time and effort, and direct financial costs (Morse, Barrett, Mayan, Olson, & Spiers, 2002; Venkatesh et al., 2013). Using an ICR method can reduce bias in coding, as it allows coders to discuss any discrepancies that they detect in their content analysis (Gaskin, Berente, Lyytinen, & Yoo, 2014). "High reliability makes it less likely that bad managerial decisions will result from using the data" (Rust & Cooil, 1994, p. 11).

Different terminology has been used to describe ICR. There is considerable confusion between the terms "inter-rater reliability" (IRR) and "inter-coder reliability" (ICR). IRR refers to a situation where two or more raters independently assign a value to each of the items they assess and then check how similar or different their ratings are (Gwet, 2014; Hallgren, 2014). For example, if two teachers independently evaluate ten student projects by giving a quality score from one to five to each project and then compare how similar or different their ratings are, the practice is called IRR check. On the other hand, ICR refers to the situation where coders in a content analysis activity independently relate pre-defined qualitative codes (e.g. accuracy, fitness for purpose, and availability as dimensions of quality of information) to related *data chunks in a transcript* and the level of agreement among coders (similarity between the coders' coding sheet) is measured (Campbell, Quincy, Osserman, & Pedersen, 2013; MacPhail, Khoza, Abler, & Ranganathan, 2016). The fact that ICR is used for evaluating reliability of content analysis means that the process of checking ICR is often more complex than the process of checking IRR. ICR often involves situations where the data could be of any type (e.g. nominal, ordinal, and ratio), and there are often many codes in the transcripts, typically more than the number of categories or values that raters use for IRR check. This requires developing a coding scheme in a way that facilitates analysis of the ICR evaluation, and requires a more careful training of coders on how to use the coding scheme. The confusion between IRR and ICR is widespread. Interestingly, even our main sources (e.g. methodology papers such as Hayes & Krippendorff, 2007 and De Swert, 2012), which have been specifically written about an ICR method, have provided examples of using the method for checking IRR. In this study, we focus exclusively on ICR.

In this paper, we aim to harmonize and systematize methodological advice on selecting and using ICR methods and provide guidelines for qualitative researchers in the information management field. We conducted an extensive review of literature to identify the characteristics of the ICR methods available and to propose our framework. First, we used these characteristics as criteria to review and compare the methods. The purpose is to enable researchers to select the most suitable method that meets the specific characteristics of a research study. We also present a snapshot of the current status of information management literature in terms of using these methods. Second, we propose an approach for applying an ICR method for the content analysis for a qualitative information management study, supported by an illustrative example. The paper ends with the discussion and conclusion sections.

2. The process of literature review

We reviewed conference and journal papers (regardless of their rank and date of publication) and research methodology books in information management, business, health, psychology, education, and other fields related to broad social sciences research (e.g. communication and media studies). The keywords for searching literature included: "inter-coder reliability", "inter-coder agreement", and "inter-coder check". We used ten databases, including: AISeL, SpringerLink, ScienceDirect, EBSCOhost databases, ABI/INFORM collection (via ProQuest), ACM Digital Library, Emerald Insight, Informs PubsOnline, Taylor & Francis Online, and Wiley Online Library. We conducted our literature review progressively from January 2019 to March 2020.

We categorized the papers into two types: 1) Studies which examined one or more of these methods, or provided a detailed description or critique of them; and 2) studies which presented a process for using these methods. Based on this selection criteria, our literature search and refinement included two rounds. The first round was using the advanced search feature of databases to identify studies that included one or more of the keywords in their titles and abstracts. This resulted in identifying 30 studies. In the second round, these papers were further refined by reviewing their full text. This reduced the number of relevant studies to fourteen. We then checked the forward and backward citations, regardless of date of publication, using Google Scholar to check the comprehensiveness of the results of our overall search process. This final round did not lead to identifying any new studies. Interestingly, none of the studies selected are in the Information Management field. They are predominantly from communication and media, sociology, health and education fields of research, with communication and media being the primary field.

Papers which have specifically provided a detailed description or critique of one or more ICR methods include: (2014a, 2014b), Hayes and Krippendorff (2007), Lombard, Snyder-Duch and Bracken (2002), Olson et al. (2016), Stevens, Lyles and Berke (2014), and Zhao, Liu and Deng (2013). Based on the insights we gained from reviewing this set of papers, we identified the characteristics of the ICR methods and used them for developing a set of criteria that researchers need to consider when selecting the most suitable ICR method for their study (RQ1; section 3 and Appendix A).

Papers that present a process for evaluating ICR include: Burla et al. (2008), Compton, Love, and Sell (2012), Campbell et al. (2013), Hruschka et al. (2004), Kurasaki (2000), MacPhail et al. (2016), and Ruggeri, Gizelis, and Dorussen (2011). We used these papers for developing our process for selecting and using a suitable ICR method (RQ2; section 5).

3. Characteristics of ICR methods and a comparative review

Based on our literature analysis we identified the following characteristics that researchers need to consider when they want to select an ICR method for their content analysis:

1 *Type of data:* the type(s) of data (nominal, ordinal, interval and ratio) the method is applicable for (Feng, 2014a, 2014b; Hayes & Krippendorff, 2007; Lombard, Snyder-Duch, & Bracken, 2002; Zhao, Liu, & Deng, 2013).
2 *Number of coders:* whether the method can be applied where more than two independent coders are involved. Content analysis of a study in which a high level of risk and sensitivity is involved[3] in its

[3] Presenting criteria for assessing the level of risk or sensitivity of a study is out of the scope of this paper. We suggest researchers assess the level of risk involved in their specific project topic through team discussions, consultations

findings may require more than two independent coders (Feng, 2014a, 2014b; Hayes & Krippendorff, 2007; Lombard et al., 2002; Olson et al., 2016; Stevens, Lyles, & Berke, 2014). Examples include studies in the area of digital health where a high level of medical risk and human ethics are involved (e.g. a study on health practitioners' views about a patient management system, where patients' health and safety can be directly affected by the findings of the study) (Cypress, 2017); and a project management study in which significant financial cost will be incurred if a wrong decision is made based on the findings (Birolini, 2012).

3 *Missing codes:* whether the method allows ICR evaluation where there are missing codes. This is important because sometimes coders may omit one or more codes in reporting the results of their content analysis (Feng, 2014a, 2014b; Hayes & Krippendorff, 2007; Zhao et al., 2013).

4 *Significance of chance in agreement:* whether the method calculates or minimizes the role of 'chance' in the independent coders' agreement on a code. Similar to the second point above, this is particularly important for a study in which a high level of risk and ethics is involved (Feng, 2014a, 2014b; Hayes & Krippendorff, 2007; Zhao et al., 2013).

5 *General agreement on the significance of a numeric result:* an ICR method is expected to produce a coefficient (a numeric result on a probability or percentage scale), where 1.000 or 100 % shows perfect agreement, and 0.000 shows complete disagreement among coders. Researchers need to pay attention to whether there is a general agreement on the significance of the result that an ICR method produces (e.g. is 0.85 a significant result for a specific method?) (Hayes & Krippendorff, 2007; Zhao et al., 2013).

The ICR methods which have been widely accepted and used include: *Percent Agreement*, Holsti's *CR* (Holsti, 1969), Bennett, Alpert and Goldstein's (1954) *S*, Scott's *pi* (π; Scott, 1955), Cohen's *kappa* (Cohen, 1960), Fleiss's *K* (Fleiss, 1971), Gwet (2014), and Krippendorff's *alpha* (1970, 2004). Table 1 provides a comparison of these methods using the five characteristics that we have identified. Appendix A presents more detailed information about each of these methods.

As the table shows, Percent Agreement is the least flexible method and Krippendorff's alpha is the most flexible method, as it can be used with more than two coders, for any type of data and missing data. This also minimizes the effect of chance in agreement. Choosing a suitable ICR method for a study requires assessing the method based on its properties and the characteristics of the content analysis of the study. For example, consider a study which needs two coders to code nominal data, there is no missing data and the level of risk and ethical concerns have been assessed as low. Percent Agreement can be suitable for such a study, as the study does not require a method that allows analysis of any type of data by more than two coders with a very low level of chance in agreement on a code. Overall, whichever method is employed, researchers should briefly explain why the characteristics of the method are appropriate for the specific characteristics of their study.

The table harmonizes and integrates the limitations and strengths of each method, but it is also important to note that: first, it is frequently the case that the 'stronger' a method is (i.e. the more accurate the calculation of ICR is), the more complex its formula. We have presented the formula for each of the eight methods in Appendix B. When we compare the formulas in Appendix B and the level of reduction in the role of chance in agreement (Table 1), we can see the direct relationship between the level of complexity of an ICR method and the level of accuracy of its result. Compared with the simpler and less accurate ICR methods, Krippendorff's alpha, Gwet's method, and Fleiss's *K* all have complex formulas that could be hard or time consuming to use for a non-specialist user. However, widespread support by statistical software packages effectively mitigates this disadvantage. Some software and applications such as SPSS, SAS, PRAM, R, Python and AgreeStat allow calculations of all or a majority of ICR methods.[4] Second, Fleiss's *K*, Cohen's kappa and Krippendorff's alpha are all sensitive to the number of codes. In practice, Krippendorff's alpha may decrease with the increase in the number of codes, Cohen's kappa and Fleiss's *K* may increase as the number of codes increases (Lombard et al., 2002; Zhao et al., 2013), and Fleiss's *K* cannot be calculated via software for a small number of codes (e.g. currently, SPSS does not calculate it if there are ten or fewer codes).

4. Current practice

We also evaluated current practice in evaluating ICR in major journals in our field. Appendix C presents a "snapshot" of the approach used for evaluating ICR in recent studies. It explains our choices of academic journals, how we reviewed the papers, and then assesses the papers in terms of what method they selected for evaluating ICR. We found that among the papers that have been published in MIS Quarterly (MISQ) and Information Systems Research (ISR), Cohen's kappa has been used most frequently, Percent Agreement has been the second most popular method, and Krippendorff's alpha is the third most popular method. Among the papers that have been published in the International Journal of Information Management (IJIM), Percent Agreement has been the most frequently used, with Cohen's kappa second, and Krippendorff's alpha has been used by only one paper. Finally, Fleiss' *K* has been used only by one study that has been published in MISQ and by one study that has been published in IJIM, and Gwet's method has been used only by one study that is published in ISR and one study that is published in IJIM. None of the other ICR methods have been used in these studies. We also note that the justification for the ICR method selected is typically weak or absent in these studies.

Finally, we used Google Scholar to conduct a forward search for each of the ICR methods in recent literature, including journal and conference papers, books, and dissertations, which have been published from January 2019 to December 2019. Appendix D presents the findings, showing that the approach can be useful for a wide range of audiences, in addition to the information management field, particularly: health (physical and mental health), social psychology, education, communication and media, and business (including all areas such as management, marketing, and finance).

5. An approach and illustrative example for selecting and using an ICR method

In this section, we explain how we designed our approach of selecting and using an ICR method. We present the approach with an illustrative example.

5.1. Methodology for designing the approach

Our methodology involves using prior literature and synthesizing their suggestions into practical guidelines for researchers. We broadly

(*footnote continued*)
with their organization or clients, and using relevant frameworks, such as Bennett et al. (1954) and The Belmont Report (1978) that explains the ethical principles and guidelines for the protection of human subjects in health-related research studies.

[4] SPSS and SAS do not calculate Krippendorff's alpha directly, and require installing a macro which was developed by Hayes (2005), Hayes, (2009). Interested researchers can download the macro by looking for KALPHA.sps from http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html. Also, the macro for calculating Fleiss *K via* SPSS can be downloaded from https://www.ibm.com/developerworks/community/files/app#/file/48234a16-fb14−4bee-8c18−570319c57108.

Table 1
A Comparison of ICR Methods.

Method	Year	Type of data	Number of coders	Missing codes?	How well the effect of 'chance in agreement' has been considered in design of the method***	General agreement on the significance of a numeric result
Percent Agreement	Unknown	Nominal	Two	No	*No consideration* of chance in calculating the ICR coefficient	No
Holsti's CR	1969	Nominal	Two	No	*Minimal consideration* of chance in calculating the ICR coefficient	No
Bennett et al.'s S	1954	Nominal	Two	No	*Reduces* the effect of chance by the logic: the likelihood that coders randomly assign an item to the same code is based on the number of codes employed.****	No
Scott's Pi	1955	Nominal	Two*	No	*Reduces* the effect of chance to a limited degree by making the assumption that coders have the same distribution of responses.	No, but Scott (1955) and Lombard et al., 2002, 2017) suggest 0.41 − 60 for moderate agreement, 0.61 − 0.80 for substantial agreement, and 0.81 − 100 almost perfect.
Cohen's Kappa	1960	Nominal	Two**	No	*Reduces* the effect of chance by combining the strengths of Bennett et al.'s S and Scott's Pi.	No, but some (e.g. Landis & Koch, 1977) suggest 0.40 − 0.59 for moderate agreement, 0.60 − 0.79 for substantial agreement, and 0.80 − 0.99 for nearly perfect agreement.
Fleiss's K	1971	Nominal	Multiple	No	*Minimizes* the effect of chance to an optimum level by measuring the degree of agreement on codes over that which would be expected to occur by chance.*****	No, but the Automotive Industry Action Group suggest that an amount over 0.90 is preferred, and between 0.75 and 0.90 is acceptable (see the report by MiniTab Inc., 2019).
Gwet	2014	Nominal AC1, and all four types for AC2	Multiple	Yes	The latest version (Gwet, 2014) *minimizes* the effect of chance to an optimum level by benefiting from the strengths of several statistical methods (see Appendix A).	No (mainly due to its newness)
Krippendorff's alpha	1970, 2007	All four types	Multiple	Yes	*Minimizes* the effect of chance to an optimum level by benefiting from the strengths of several statistical methods (see Appendix A).*****	Yes. An outcome over 0.9 is always acceptable, between 0.8 and 0.9 is considered suitable or fair reliability, and over 0.7 is tolerable for exploratory studies.

* An extended version by Siegel and Castellan (1988) accommodates multiple coders.
** An extended version by Conger (1980) accommodates multiple coders.
***It is not possible to mention how much the effect of chance in agreement has been minimized for each method. The table therefore does not provide details on how precisely chance is reduced in each calculation.
**** See a related work by Rust and Cooil (1994), who model the loss by considering wrong judgments.
***** Zapf, Castell, Morawietz, and Karch (2016) conducted a large simulation study and identified that Fleiss' K and Krippendorff's alpha are as reliable as each other in terms of reducing the role of chance in agreement.

follow a design science paradigm for designing a research method artifact (Venable & Baskerville, 2012), considering our "approach for selecting and using an ICR method" to be an artifact. This approach to developing our artifact is similar to Nili, Tate, and Johnstone (2017), who developed a method for the analysis of focus group data using a design science approach and is consistent with Venable and Baskerville's (2012) and Gregor and Hevner (2013) guidelines. "Design Science is an appropriate paradigm for research into research methods… Applying a [Design Science Research] DSR perspective to research methods should yield increased utility in the application of research methods, better guidance in applying them and greater confidence in achieving the desired outcomes of applying them" (Venable and Baskerville, 2012, p. 399). The process of designing our approach started with analyzing and synthesizing the content of the seven papers we identified that presented a process for using ICR methods to determine the steps involved in selecting and using an ICR method. Among these papers, Kurasaki (2000) suggested an approach that comprises four steps: train and calibrate coders, code data, check the process of data analysis at a mid-point, and calculate the ICR coefficient. Burla et al. (2008) and MacPhail et al. (2016) suggested a high-level process that includes three steps: develop a coding scheme, evaluate the ICR, and conduct a final review of codes. Hruschka et al. (2004) and Campbell et al. (2013) provided some suggestions on resolving potential problems with assessing ICR. Compton, Love, and Sell (2012) suggested that the development of a coding scheme and training of coders can be labeled as the two steps of pre-testing in the overall process. We reviewed each of these papers in detail and synthesized their suggestions.

The result of our synthesis is a process that comprises five steps: (1) selecting an ICR method, (2) developing a coding scheme, (3) selecting and training independent coders, (4) calculating the ICR coefficient (which may lead to continuing the training session and iteratively coding the entire dataset), and (5) reporting the process of evaluating ICR along with the result. We provide the details on each of these steps in the next section.

Overall, the approaches that the seven papers propose are generally incommensurate with each other, however, they are similar in one respect: all these approaches are very high-level. Compared with these resources, our proposed approach includes more specific steps and we provide specific guidance particularly on what characteristics should be considered when selecting an ICR method for a study, how to develop a coding schema that supports the calculation of the ICR coefficient, and specific guidance on reporting the process of ICR evaluation.

The approach was iteratively assessed via ongoing discussions among the research team, peer review, and expert feedback that we received on the preliminary and final versions of the approach. This feedback and review did not result in any significant revision to the approach. To illustrate, we evaluated the ICR of a research project that we conducted recently, focusing on the steps and their sequence in our approach (Fig. 1).

5.2. Our proposed approach and an illustrative example

The project aimed to identify the factors that contribute to user persistence in solving their own self-service technology problems. We focused on studying why and how people use information to solve their problems with self-service IT in the workplaces, such as problems with a research grant application system or problems with a self-service financial reconciliation system for corporate credit card users. Data were gathered through qualitative individual interviews with 30 users who had experienced such problems using internal IT systems at a large tertiary institution. Each interview took between 20 − 45 min and was fully transcribed. The transcripts included over 200 pages. We employed three independent coders who were using a coding sheet that included 33 codes (see Appendix E which presents a part of the coding sheet). The type of data was nominal and there were no missing codes. The project was assessed as low risk by the researchers and by their institution's research ethics committee. Below, we describe our project activities in relation to each step of our proposed approach for ICR evaluation.

5.2.1. Select an ICR method

In this step, the nature of the data and coding scheme, number of coders, and the need for minimizing the effect of chance in agreement are considered, and the appropriate method is selected. For this step, Table 1 provides a comparison of the methods based on the five evaluation criteria that emerged from our literature synthesis. Researchers can apply these criteria to the specific characteristics of their research project to determine the most appropriate method. If researchers identify more than one suitable method for their research, using all or at least two of those methods can enhance the robustness of their ICR evaluation.

Example: Based on the type of data (nominal); the number of independent coders (three); the completeness of the data (no missing codes); and because the risks and ethical concerns related to the project were assessed as negligible by the research team and the research ethics committee at our institute, Krippendorff's alpha, Fleiss's *K*, and Gwet's method were identified as suitable ICR methods for the research project. We used all three methods.

5.2.2. Develop coding scheme

Developing a coding scheme may be done 'a priori' (i.e. developing the codes based on available knowledge and review of literature), or inductively and iteratively from data. However, eventually, the coding scheme will be finalized and the definition of each code established. At the minimum, the coding scheme needs to include the codes and their definitions. Coding rules and relevant direct quote from the transcript could also be included in the coding scheme (Burla et al., 2008; MacPhail et al., 2016).

Example: An a priori coding scheme was developed in the form of a table (Appendix E) by the research team based on an extensive literature review, and by following the advice by well-established qualitative resources including Miles, Huberman, and Saldana (2014) and King and

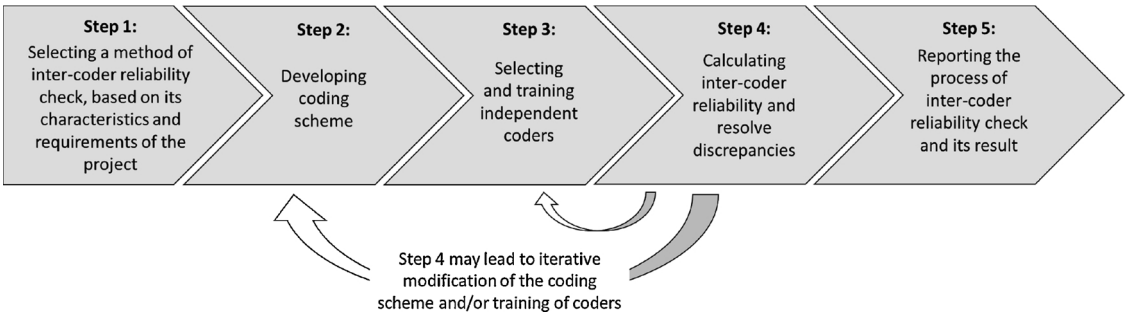

Fig. 1. An Approach for Selecting and Using an ICR Method.

Cassell (1998). As the result of initial content analysis of the interviews, the research team identified 33 codes. The first column of the coding scheme included a number for each code, the second column included the labels of the codes (e.g. 'system interactivity', 'IT self-efficacy', 'expected time' and 'expected effort'), and the third column included a specific definition for each code. We asked each coder to consider the definitions of the codes and independently allocate the number of a code that matches the meaning of a related data chunk in the transcripts, and then assessed agreement between the coders based on the level of similarity between the numbers that they assigned to the same data chunks. Coders could be given flexibility in the way they report the findings (e.g. paper and pencil or spreadsheets). If spreadsheets have been used, the results can more efficiently be exported to software that calculates the ICR coefficient.

5.2.3. Select and train independent coders

Ideally, the selected coders have domain-specific knowledge and some qualitative coding experience but no previous experience with the project (Compton et al., 2012; MacPhail et al., 2016). Coders need to be fully conversant with the definitions of the codes and confident in using the coding sheet. We suggest that all coders use the coding sheet to code randomly selected text from the transcripts. The training may be continued until all coders feel confident in using the coding sheet.

Example: three coders were selected for this project. The original research team explained the purpose of the research and conducted a coding practice with a small sample of the interview transcripts (10 pages out of the 200 pages) and asked each coder to independently allocate the number of each code to its related data chunk. When the coders were feeling confident about this process, they were asked to code half of the data independently. An initial screening of the coding sheet showed that coders were feeling confident about this activity. Therefore, they were asked to use the coding sheet to code the rest of the transcripts and record the coding results.

5.2.4. Calculate ICR and resolve discrepancies

In this step, the selected ICR method (or methods) is used. Ideally, a software package is employed to support calculating the ICR coefficient (section 3). A weak result could lead to further training of the coders and iterative coding of the dataset. We note that many of the papers that we reviewed in Appendix C stopped their process of checking ICR once their calculation showed a high result. However, even after achieving a high ICR result, if there is any discrepancy about a code, it needs to be discussed and a consensus developed. Alternatively, the discrepancy needs to be briefly discussed in the final report.

Example: Having selected Krippendorff's alpha, Fleiss's *K*, and Gwet's method as the most appropriate methods, in order to calculate the ICR coefficient, we entered the number for each data chunk into MS Excel (see the left column in Table 2 representing a portion of the data). Next, the code number that each coder allocated to each data chunk was entered in the second, the third and the fourth columns for that data chunk (see section 5.2.2 for information on allocating a code number to its related data chunk). We note that this way of allocating numbers for data chunks and codes to check inter-*coder* reliability is completely different with the studies (e.g. De Swert, 2012 and Hayes & Krippendorff, 2007) which have used ICR methods to check IRR. We imported the Excel sheet to SPSS to calculate ICR through the Krippendorff's alpha and Fleiss's *K* methods and imported it to AgreeStat to calculate ICR using Gwet's method. The results from Gwet's method, Fleiss's *K,* and Krippendorff's alpha were 0.82, 0.81, and 0.81 respectively. These results are considered to be a suitably high level of agreement (Krippendorff, 2004). The discrepancies were then discussed by the three coders in a separate session and complete agreement was achieved.

Table 2
A Portion of the Data on an Excel Sheet.

DataChunk	Coder1	Coder2	Coder3
1	3	3	3
2	2	2	2
3	6	6	6
4	12	12	10
5	16	16	16
6	21	21	21
7	8	8	8
8	7	7	7
9	9	9	9
10	4	10	10

As discussed, because the ICR check was done by more than two coders, other methods such as Percentage Agreement are not applicable. However, if we conduct an ICR check for only two coders, in this case Coder 1 and Coder 2, the result of the ICR check is 0.83 (Percentage Agreement), 0.78 (Gwet's method), and 0.77 (both Krippendorff's alpha and Fleiss's K). Such differences can be particularly important for high risk projects, which require a high reliability of findings (see section 3).

5.2.5. Report the process of checking ICR and its result

Finally, the overall process of checking ICR and its result should be communicated when reporting on the project. A concise explanation of how the approach was conducted can provide convincing information for those who seek rigor and trustworthiness in research. Appendix F presents a general structure for reporting the process of ICR check and its result.

Example: In a recent description of our research project (withheld for review), we presented a paragraph similar to the text in Appendix F to briefly explain how we conducted our five-step approach.

6. Discussion

Evaluating ICR is a key quality metric for the credibility and trustworthiness of content analysis in many qualitative information management studies. However, available guidance is diffuse, and frequently obtained from publications in other disciplines. Although high quality publication outlets expect the use of an ICR method for qualitative studies, our brief overview of recent practice in leading information management journals suggests the practice is not consistent or rigorous. Many studies have *not* provided a discussion of the method they used for evaluating ICR, and this absence underscores the importance of our study.

We clarified terminology, noting that ICR and IRR are not synonymous and interchangeable, although they are often confused. Our contribution is to synthesize, harmonize and systematize methodological advice about approaches for evaluating ICR, and present guidance that aims to help qualitative researchers to improve their practice of evaluating ICR for their research studies. By comparing the range of methods available, we show that unsurprisingly, the accuracy and flexibility of the calculation of ICR increases with the statistical complexity of the method. However, as with all statistical tests, used in research contexts, the selection of the most appropriate method should be an informed decision based on the criteria we describe, including the project requirements, the characteristics of the data, and the number of coders. The challenges of complexity are greatly reduced by the availability of functions to calculate ICR metrics in leading statistical software packages. However, the ease with which an ICR metric can be obtained 'at the touch of a button' easily lead to an uncritical approach to selecting and employing an ICR method.

We aim to "lift the hood" on ICR calculation methods that are often "black-boxed" by software packages. We offer qualitative researchers an accessible description of the various methods, detailed evaluation criteria, and a process for selecting and applying the most appropriate method depending on the characteristics of their study. Our proposed

approach was accompanied by an illustrative example of how it can be used.

While we identified Krippendorff's alpha as an ICR method which is likely to be relevant in many contexts, we recommend that researchers not simply select this method as a default. In all cases, researchers need to explain why the characteristics of their selected method are appropriate for the specific characteristics of their study.

Our study is not without limitations. We focus primarily on the methods themselves, with only a limited snapshot of current practice in their use (Appendix C). We did not carry out a full literature analysis of the current state of practice with regard to calculation of ICR. Future research could carry out a more comprehensive evaluation of literature, possibly using bibliometric analysis, to examine how authors have approached the selection, execution, and reporting of ICR methods, including recommendations for improvement.

7. Conclusion

Evaluating inter-coder reliability is becoming standard practice for qualitative studies, yet the guidance about that practice has been scattered and lacking in detail. Our framework contributes to the growing trend of providing focused methodological sources in the information management field, and can be useful for both novice and experienced information management researchers in evaluating inter-coder reliability of their content analysis. We suggest carrying out an inter-coder reliability check for any content analysis wherever possible. Moreover, our forward search of recent academic studies shows that our proposed approach can be useful for qualitative content analysis in several other research fields as well, which should not be surprising, as information management is a multi-disciplinary field and it has grown to the extent that our methodological contributions are expected to be useful for our reference disciplines, as well.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Alireza Nili: Conceptualization, Methodology, Validation, Writing - original draft. **Mary Tate:** Visualization, Writing - review & editing. **Alistair Barros:** Resources, Writing - review & editing. **David Johnstone:** Writing - review & editing.

Appendix A. Table A.1. A Review of ICR Methods

Method	Description of the method based on the five characteristics
Percent Agreement	Percent Agreement is the simplest ICR method (Feng, 2014a, 2014b). It focuses on the number of codes which have been considered for their right (or related) data chunks by independent coders. The method can be used only by two coders and only for nominal data (Zhao et al., 2013). Percent Agreement does not account for agreement that could occur by chance (the two coders may agree on some of the codes by chance); therefore, it may overestimate true agreement between coders (Feng, 2014a, 2014b; Krippendorff, 2013; Zhao et al., 2013). The method is often suitable for a low risk study that does not require a high level of precision in assessing ICR (e.g. a study that the research team or their organization assesses it as a low risk study to people's health or does not lead to a significant financial loss). As the number of codes increase, however, high agreement between coders becomes more difficult, allowing for argument on the suitability of Percent Agreement for high risk projects.
Holsti's CR	Holsti's (1969) method is a variation of the Percent Agreement and can be used by two coders only. "[It] accounts for situations in which the coders evaluate different units. The result is often calculated not for a single variable but across a set of variables, a very poor practice which can hide variables with unacceptably low levels of reliability" (Lombard et al., 2002, p. 591). We however note that similar to the Percent Agreement method, as the number of codes increase, high agreement between coders becomes more difficult, which allows for argument on the suitability of Holsti's CR, particularly for a low risk study.
Bennett et al.'s S	Similar to the Percent Agreement, Bennett et al.'s (1954) S is limited to two coders and to nominal data. Compared with Percent Agreement, however, it is considered 'more' reliable and less likely to be the influenced by chance (Hayes & Krippendorff, 2007; Lombard et al., 2002). The method reduces the effect of chance by equaling the ratio of observed non-chance-agreement to possible non-chance-agreement. It can be used regardless of complexity of the content analysis (e.g. the number of codes does not affect calculation of ICR significantly). However, "S is inflated by the number of unused categories that the author of the instrument had imagined and by rarely used categories in the data" (Hayes & Krippendorff, 2007, p. 80). In the past three decades, the method has been revised several times as the ICR coefficient Ir. (Perreault & Leigh, 1989).
Scott's pi	Scott's pi (π; Scott, 1955) is a very similar method to the Bennett et al.'s S. Scott's pi can be used only by two coders and for nominal data. It accounts for the number (e.g. few or many) of codes and their distribution in the transcript, allowing the method to correct Percent Agreement by taking into account the agreement that can occur among coders by chance. It equals the ratio of observed non-chance agreement to possible non-chance agreement to identify how often the coders agreed when they were not guessing (Hughes & Garrett, 1990; Lombard et al., 2002; Zhao et al., 2013). Later, the method was slightly revised by Siegel and Castellan (1988) who extended the method to accommodate multiple coders.
Cohen's kappa	The Kappa (κ) coefficient (Cohen, 1960) was proposed as an alternative to π. The method corrects the Percent Agreement method, just as do π and *S*. The most important issue with the method is the difficulty in interpreting its result. There is not just a single threshold to indicate what a high, acceptable, or low agreement is (Olson et al., 2016; Zhao et al., 2013). However, some methodologists (often, the researchers who have cited Landis & Koch, 1977) suggest that 0.60 is the threshold for "substantial agreement" and 0.80 is the "nearly perfect" agreement. A change in the number of codes can also influence the result. Kappas become higher as number of codes increases, making it even less likely for researchers to be able to clearly mention how significant the magnitude of the result is (Lombard et al., 2002; Zhao et al., 2013). Later, Conger (1980) extended κ to accommodate multiple coders; however, it is still limited to nominal data. We also note that Rogot and Goldberg (1966) A_2 coefficient is equivalent to the Cohen's kappa method.
Fleiss's *K*	Fleiss (1971) generalized Scott's pi to any number of coders and called it kappa, which was renamed *K* by Siegel and Castellan (1988). *K* expresses the extent to which the observed number of agreements among coders exceeds what could be expected if all coders made their coding in a completely random way. In other words, it measures the degree of agreement on codes over that which would be expected to occur by chance (Feng, 2014a, 2014b; Olson et al., 2016). It is a reliable method in terms of considering chance in agreement. It is however limited to nominal data. The outcome of the method ranges from 0 (no agreement at all) to 1 (perfect agreement) (Hayes & Krippendorff, 2007; Olson et al., 2016).
Gwet (2014)	In a more recent work, Gwet (2014), revised and generalized Bennett et al.'s S, Scott's pi, Fleiss's *K* and Cohen's Kappa to accommodate multiple coders and missing codes that may occur as the result of mistake in coding of data by a coder. It should be noted that these methods had gone through several revisions, and multiple equivalent versions of them had been proposed in several other forms such as Guilford's G (Holley & Guilford, 1964); Brennan and Prediger (1981) free marginal kappa coefficient; Byrt, Bishop, and Carlin (1993) prevalence and bias adjusted kappa coefficient; Janson and Vegelius (1979) C score; Maxwell (1977) random error coefficient; and Potter and Levine-Donnerstein (1999) redefined pi coefficient. We note that a few previous versions of Gwet (2002); 2008; 2010) method (sometimes called AC1 and AC2) also exist, but they have not been adopted significantly by researchers yet (Feng, 2014b). The AC_1 version of Gwet's method and its weighted version (known as AC_2) are extensively discussed in Gwet (2014). AC1 is simply AC2 restricted to nominal (categorical) data.

Krippendorff's alpha	Krippendorff's alpha (1970; 2004) resolves many of the limitations of the other ICR methods. The method measures agreements for nominal, ordinal, interval, and ratio data, and also allows for measuring reliability with missing codes. The method can also be used where more than two coders are coding the data, and minimizes the effect of chance in agreements on the codes (Feng, 2014a, 2014b; Park & Park, 2015). These characteristics of the method are mainly because it embraces several known reliability coefficients, including Scott's π for its two-coder nominal data calculation; Pearson, Lee, Warren, Fry, and Fawcett (1901) intraclass-correlation coefficient for its two-coder interval data calculation; a form of Spearman's rank correlation coefficient ρ for its two-coder ordinal data calculation; and (Krippendorff, 1970) for its extension to more than two coders. Perfect reliability is expressed by 1.000, and 0.000 shows the absence of any degree of reliability. The more coders independently code the data, the more difficult it is to achieve a high level of agreement among coders; therefore, this numerical measure may decrease with the increase in the number of coders. There is a general agreement that any outcome over 0.9 is always acceptable, over 0.8 is considered 'suitable', and over 0.7 is tolerable for an exploratory study (Feng, 2014a, 2014b; Gerdes, Stringam, & Brookshire, 2008; Lombard et al., 2002). Krippendorff (2012) suggests that a result that is over 0.8 guarantees fair reliability, and any result between 0.667 and 0.8 could support tentative findings of the content analysis.

Methods which Should Not Be Confused with ICR Methods.

Interclass Correlation Coefficients, Cronbach's alpha, Chi-square, and Pearson's r: These are the methods which should not be confused with ICR methods. "Because interclass correlation coefficients do not consider systematic coding errors by judges, they are in-adequate to assess inter-coder reliability" (Hughes & Garrett, 1990, p. 187). Similarly, chi-square (which measures association), Cronbach's alpha (or αC which measures internal consistency; Cronbach, 1951), and Pearson's r (which measures correlation) are not ICR methods, as none of these methods measures the degree of agreement among coders. For example, although αC is called a reliability coefficient, it does not specifically measure agreement between coders. In fact, it is a statistic for interval or ratio level data that focuses on the consistency of judges when numerical judgments are required for a set of units. It calculates the consistency by which people judge units without any aim to consider how much they agree on the units in their judgments. In addition to these four measures, *rwg* (Brown & Hauenstein, 2005; James, Demaree, & Wolf, 1984) "is a frequently used index for interrater agreement on Likert-type scales" (Dong, Fang, & Straub, 2017), therefore we do not consider this method as an ICR method.

Appendix B

Equation B.1: Percent agreement

$$\text{Percent agreement} = {}^{A}\!/_{N}$$

where

A is the number of codes (i.e. codes in the coding sheet) on which both coders are in agreement

N is the total number of codes in the coding sheet

Equation B.2: Holsti's C.R. (Holsti, 1969**)**

$$\text{C.R.} = \frac{2\text{M}}{\text{N}_1 + \text{N}_2}$$

where

M is the number of codes (i.e. codes in the coding sheet) on which both coders are in agreement

N1 is the number of codes coder 1 reported

N2 is the number of codes coder 2 reported

Equation B.3: Bennett et al.'s S (Bennett, Alpert, & Goldstein, 1954**)**

$$S = \left(\frac{K}{K-1}\right)\left(a_o - \frac{1}{K}\right)$$

where

$$a_o = {}^{A}\!/_{K}$$

a_o is observed agreement

A is the number of codes (i.e. codes in the coding sheet) on which both coders are in agreement

K is the total number of codes in the coding sheet

Equation B.4: Scott's pi (π; Scott, 1955**)**

$$\pi = \frac{p_o - p_c}{1 - p_c}$$

where

$$p_o = \left(\frac{n_{ii} + n_{jj}}{n}\right)$$

$$P_c = \left(\frac{m_i}{n}\right)\left(\frac{m_i}{n}\right) + \left(\frac{m_j}{n}\right)\left(\frac{m_j}{n}\right)$$

$$m_1 = \frac{n_{+i} + n_{i+}}{2}$$

$$m_2 = \frac{n_{+j} + n_{j+}}{2}$$

p_o is the proportion of agreement between coders

P_c is the proportion of expected agreement by chance

$n_{ii}$is the number of items (i.e. data chunks) both coders assigned to code i
$n_{jj}$is the number of items both coders assigned to code j
n is the total number of items
$n_{i+}$is the number of items coder 1 assigned to code i
$n_{j+}$is the number of items coder 1 assigned to code j
$n_{+i}$is the number of items coder 2 assigned to code i
$n_{+j}$is the number of items coder 2 assigned to code j

Equation B.5: Cohen's kappa (Cohen, 1960**)**

$$K = \left(\frac{p_o - p_c}{1 - p_c}\right)$$

where

$$p_o = \frac{n_{ii} + n_{jj}}{n}$$

$$P_c = \left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+i}}{n}\right) + \left(\frac{n_{j+}}{n}\right)\left(\frac{n_{+j}}{n}\right)$$

p_o is the proportion of agreement between coders
P_c is the proportion of expected agreements by chance
$n_{ii}$is the number of items (i.e. data chunks) both coders assigned to code i
$n_{jj}$is the number of items both coders assigned to code j
n is the total number of items
$n_{i+}$is the number of items coder 1 assigned to code i
$n_{j+}$is the number of items coder 1 assigned to code j
$n_{+i}$is the number of items coder 2 assigned to code i
$n_{+j}$is the number of items coder 2 assigned to code j

Equation B.6: Fleiss' Kappa (Fleiss, 1971)

$$K = \left(\frac{p_a - p_s}{1 - p_s}\right)$$

where

$$p_a = \frac{1}{mn(m-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2 - mn\right]$$

$$p_s = \sum_{j=1}^{k} q_j^2$$

$$q_j = \frac{1}{nm}\sum_{i=1}^{n} x_{ij}$$

n = the number of items (data chunk i = 1, 2, …, n)
k = the number of codes (code j = 1, 2, …, k)
m = the number of coders for each data chunk
x_{ij} = the number of coders that assign code j to data chunk i

Equation B.7: Gwet (2014)

Gwet (2014) extensively discusses two ways of calculating ICR, including: Gwet's AC_1 and Gwet's AC_2 (weighted version of AC_1). AC1 is restricted to the nominal (categorical) data. AC2 is the extension of AC1 to ordinal, interval and ratio ratings. In addition, based on Gwet's (2014) advice, below are the generalized formulas (applicable to any type of data and number of coders) for evaluating ICR:

$$\gamma = \left(\frac{p_o - p_c}{1 - p_c}\right)$$

where

$$p_o = \frac{1}{n'}\sum_{i=1}^{n'}\sum_{k=1}^{q}\frac{r_{ik}(r_{ik}^* - 1)}{r_{ik}(r_{ik} - 1)}$$

$$r_{ik}^* = \sum_{l=i}^{q} w_{kl} r_{il}$$

$$p_c = \frac{T_w}{q(q-1)}\sum_{k,l}\pi_k(1 - \pi_k)$$

$$T_w = \sum_{k,l} w_{kl}$$

$$\pi_k = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{ik}}{r_i}$$

q is the total number of codes

w_{kl} is the weight associated with two coders assigning an item (i.e. a data chunk) to codes k and l

r_{il} is the number of coders who assigned item i to code l

n' is the number of items that were coded by two or more coders

r_{ik} is the number of coders who assigned item i to code k

r_i is the number of coders who assigned item i to any code

n is the total number of items

Equation B.8: Krippendorff's alpha (Krippendorff, 1970, 2011**)**

Dependant on the circumstances (e.g. different types of data and different number of coders in two distinct studies), calculating Krippendorff's alpha can take different forms. See Krippendorf (2011) for different ways of evaluating ICR in different circumstances. Based on Krippendorf's (2011) advice, below are the generalized formulas (applicable to any type of data and number of coders) for evaluating ICR:

$$\alpha = \left(\frac{p_o - p_c}{1 - p_c}\right)$$

where

$$p_o = p'_o(1 - \varepsilon_n) + \varepsilon_n$$

$$p'_o = \frac{1}{n'}\sum_{i=1}^{n'}\sum_{k=1}^{q}\frac{r_{ik}(r_{ik}^{*} - 1)}{r^{-}(r_i - 1)}$$

$$\varepsilon_n = \frac{1}{n'r^{-}}$$

$$r_{ik}^{*} = \sum_{l=i}^{q} w_{kl}r_{il}$$

$$r^{-} = \frac{1}{n'}\sum_{i=1}^{n'} r_i$$

$$p_c = \sum_{k,l}^{q} w_{kl}\pi_k\pi_l$$

$$\pi_k = \frac{1}{n'}\sum_{i=1}^{n'}\frac{r_{ik}}{r_i}$$

q is the total number of codes

w_{kl} is the weight associated with two coders assigning an item (i.e. a data chunk) to codes k and l

r_{il} is the number of coders who assigned item i to code l

n' is the number of items that were coded by two or more coders

r_{ik} is the number of coders who assigned item i to code k

r_i is the number of coders who assigned item i to any code

Appendix C

We provide a "snapshot" of recent top quality research at a scale that is sufficient for us to provide an insight about the current status of research studies that have (or have not) used an ICR method. We restricted our search to the papers that have been published in the last five years (from March 2013 to June 2019) in the three elite journals, including MIS Quarterly (MISQ), Information Systems Research (ISR), and International Journal of Information Management (IJIM). In recent information management research "there is greater focus on managing activities that make changes in patterns of behavior of customers, people, and organizations, and information that leads to changes in the way people use information to engage in knowledge-focused activities" (IJIM, 2019). The topic of papers that are published in these three elite journals are commensurate with this notion. Also, the journals cover diverse areas of practice (where the roles of information technology and information user are pertinent, such as business and digital marketing, digital health, education, and digital government) which suit the nature of information management research.

Because it was not possible to use the common methods of literature search (e.g. keyword search based on title, abstract, and keywords) to reliably identify and review the papers that have (or have not) used an ICR method, we downloaded and manually reviewed all papers that have been published in that period of time in a journal to provide the "snapshot" of the literature. Because of the nature and applicability of ICR methods, we only considered qualitative and mixed-method papers which included a qualitative coding (content analysis) activity.

Table C.1. The Use of ICR Methods by the Studies that Have Been Published from January 2013 to March 2019 in MISQ, ISR, and IJIM.

Method	Number of MISQ papers that have used the method	Number of ISR papers that have used the method	Number of IJIM papers that have used the method
Cohen's Kappa	12 papers (e.g. Tan, Benbasat, & Cenfetelli, 2016, and Tsai & Bagozzi, 2014)	9 papers (e.g. Breward, Hassanein, & Head, 2017, and Lindberg, Berente, Gaskin, & Lyytinen, 2016)	8 papers (e.g. Hsieh & Hsieh, 2013, and Li, Zhang, Tian, & Wang, 2018)

Percent Agreement	10 papers (e.g. Karhade, Shaw, & Subramanyam, 2015, and Ou et al., 2013)	4 papers (e.g. Bauer, Franke, & Tuertscher, 2016, and Ruckman et al., 2015)	12 papers (e.g. Mäntymäki & Salo, 2015, and Cheng, Fu, & de Vreede, 2017)
Krippendorff's alpha	3 papers, including: Ludwig et al. (2014); Beck, Pahlke, and Seebach (2014), and Oh, Agrawal, and Rao (2013)	2 papers, including: Arazy, Daxenberger, Lifshitz-Assaf, Nov, and Gurevych (2016) and Oh, Eom, and Rao (2015)	1 paper (Akter et al., 2019)
Fleiss's K	1 paper (Grover & Lyytinen, 2015)	0	1 paper (Palese & Usai, 2018)
Gwet (2014)	0	1 paper (Prabuddha et al., 2013)	1 paper (Antioco & Coussement, 2018)
Holsti's CR	0	0	0
Bennett et al.'s S	0	0	0
Scott's Pi	0	0	0

Among the 76 studies that have been published in MISQ and have performed content analysis of qualitative data, we identified only 34 papers that have used an ICR method. Among these, 30 papers have mentioned the details about the method that they have employed. Among the 49 studies that have been published in ISR and have performed content analysis of qualitative data, 26 papers have used an ICR method, among which, only 17 papers have mentioned the details about their use of the method. Finally, among the studies that have been published in IJIM and have performed content analysis of qualitative data, 31 papers have used an ICR method, among which 23 papers have mentioned the details about the method that they have employed. Table C.1 presents a summary of these results. We do not presume to comment on the overall quality of these papers. However, the absence of any discussion of this issue in some of the recent papers in our leading journal is a further motivation for the importance of our study.

Appendix D. Table D.1. Research fields in which ICR methods have been cited frequently

Method	Year of publication	Number of citations by December 2019	Research fields in which the method has been cited frequently (from January 2019 to December 2019)*
Percent Agreement	Unknown	The original source has not been mentioned in literature, hence we were unable to conduct a forward search for this method.	Anecdotally, Percent Agreement is the most frequently used method in most areas of research that requires content analysis because of its simplicity (Feng, 2014a, 2014b).
Holsti's CR	1969	9286	Business (including all areas such as management, marketing, and finance), health (physical and mental health), social psychology, information management, and education
Bennett et al.'s S	1954	232	Artificial intelligence (focusing on facial recognition, speech/voice recognition, and image processing)
Scott's Pi	1955	2308	Information management, communication and media, Health (physical and mental health), social psychology, and business (including all areas such as management, marketing, and finance)
Cohen's Kappa	1960	33,430	Health (physical and mental health), social psychology, information management, education, and business (including all areas such as management, marketing, and finance), public policy, and energy and sustainability
Fleiss's K	1971	5803	Health (physical and mental health), social psychology, information management, and business (including all areas such as management, marketing, and finance)
Gwet	2014	1479 **	Health (physical and mental health), social psychology, education, information management, communication and media, and business (including all areas such as management, marketing, and finance), public policy, and energy and sustainability
Krippendorff's alpha	1970	249 ***	Health (physical and mental health), social psychology, information management, education, business (including all areas such as management, marketing, and finance), communication and media

* We have listed these fields of research based on how often they have been mentioned for each method, from the highest to lowest frequency respectively.

** Please note that the number of citations refers to a relatively recent handbook by the author and does not include the citations received for the author's prior relevant work on evaluating ICR.

*** The work by Hayes and Krippendorff (2007) which clarifies a few points about the method has received 2759 citations by December 2019.

The limitations of our forward search include: first, our forward search did not cover any analysis of the text. As such, a research study that has cited a method might have practically used, discussed or simply mentioned it. Second, in addition to the fields of research we have mentioned in the table, we identified a few research fields (e.g. environment, energy and sustainability) which we have not reported in the table because we identified minimal number of publications (e.g. only five papers) in those fields. Finally, it is not possible for us to provide an exact number or percentage for how often each method has been mentioned in each field, as many of the academic outlets (journals, conferences, etc.) are interdisciplinary (e.g. health informatics, which covers both health and information management areas of research) and some of them (e.g. methodology books) target audience in a wide range of research fields. This however supports our claim that our framework can be useful for a wide range of audience - beyond the information management field. Where publication did not specify audience or an area of research as its primary area of focus, we read its title, abstract and the introduction section to decide what area or areas are the most suitable to be considered as the primary area of focus.

Appendix E. Table E.1. A part of the coding scheme which was used by the independent coders

#	Codes	Definition	Example
1	IT Self-Efficacy	Refers to individuals' judgment of his/her knowledge and skills (capability) to use IT in diverse situations.	"I believe I can handle these problems by myself."
2	Prior Knowledge		

		Prior knowledge (gained from the experience) of solving the same or a similar problem in the past.	"It depends on how much we know… so based on our previous experiences we decide what our course of action would be."
3	Attribution (of cause of the problem)	A user may attribute an event or the cause/reason for a problem to his/her own actions or to external factors such as other people (e.g. to the service provider) or technology (e.g. the IT is too hard to use or it has not been designed well)	"Well, the problem is not me." "But I usually think it is me. I usually just think I am ignorant. There should be a button somewhere that I have not seen or a drop down box."
4	Subjective Norm	A user's perception that most people who are important to him/her (e.g., colleagues or friends) think that he/she should or should not solve the technology problem.	"…their opinion was important; I [felt] I have to solve it."
5	Perceived Control over Solving the Problem	Users' perception of the degree to which they have control over their problem-solving behavior/activities	"…because, for my problem, I cannot have the admin right [to make the necessary changes], so sometimes I found the answer, but I could not apply it to my computer."
6	System Interactivity	User's perception of how well a system responds to commands and how easily it enables arrangement of the amount, sequence and style of information and problem-solving activities	"scanning; yeah you press scan, yeah fine, how do I get that on that university computer now? It was straightforward. It looked like, okay, transfer to external device and stuff like that. So I just followed all the things [it asked me to do], like talk me through, basically."
7	Perceived Ease of Use	The degree to which a user believes that using the system would be free of effort (the technology is easy to use).	"It was really difficult to find the option you want… Even if you find it is not easy enough."
8	Perceived Usefulness	The degree to which a user believes that using the system enhances his or her job performance, or the technology is useful for the user in general.	"For software that you want to know it completely, be in charge and see what that software does, persistence is very important…"
9	Expected Time (required for using self-help information)	User's perceived level/amount of time (required for using self-help information) to solve a problem through a method of solving IT problems.	"…, but, overall, it was taking too long and I needed the system to upload my document."
10	Expected Effort (required for using self-help information)	User's perceived level/amount of effort (required for using self-help information) while trying to solve the problem through a method of solving IT problems.	"I may be just investing more effort and some cost… You may continue to make the investment."

Appendix F

Please note that while we recommend the concise text below as the general structure for reporting the process of ICR check and its result, the text may need to be revised based on specific requirements of a research project and the publication outlet and based on flow of information in that research manuscript:

The type of data was *[T type of data]* and there [*was N number 'or' was no*] missing code. The project was assessed as [*low risk, negligible, high risk, or…*] by the researcher(s) and by their institution's research ethics committee. In order to check the inter-coder reliability of the content analysis, we followed the process of inter-coder reliability check suggested by [*a reference*]. The steps of the process included [*mention the steps*]. We selected [*Method X*] [*because…*] and used [*Software Y*] to calculate inter-coder reliability. We employed *[N number of]* independent coders. After training the coders in *[N number of]* sessions and ensuring that they feel confident in coding, the coders used the coding sheet which we provided for them to code the entire transcript. The final result was [*R result*] which is considered [*satisfactory, ideal, or…*]. [*If discrepancies between the coders existed and if the researchers conducted a session to resolve them*] we resolved the discrepancies in *[N number of]* session with the coders, achieving 100 % agreement on the codes.

References

Akter, S., Bandara, R., Hani, U., Wamba, S. F., Foropon, C., & Papadopoulos, T. (2019). Analytics-based decision-making for service systems: A qualitative study and agenda for future research. *International Journal of Information Management, 48*, 85–95.

Antioco, M., & Coussement, K. (2018). Misreading of consumer dissatisfaction in online product reviews: Writing style as a cause for bias. *International Journal of Information Management, 38*(1), 301–310.

Arazy, O., Daxenberger, J., Lifshitz-Assaf, H., Nov, O., & Gurevych, I. (2016). Turbulent stability of emergent roles: The dualistic nature of self-organizing knowledge co-production. *Information Systems Research, 27*(4), 792–812.

Bauer, J., Franke, N., & Tuertscher, P. (2016). Intellectual property norms in online communities: How user-organized intellectual property regulation supports innovation. *Information Systems Research, 27*(4), 724–750.

Beck, R., Pahlke, I., & Seebach, C. (2014). Knowledge exchange and symbolic action in social media-enabled electronic networks of practice: A multilevel perspective on knowledge seekers and contributors. *MIS Quarterly, 38*(4), 1245–1269.

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited-response questioning. *Public Opinion Quarterly, 18*(3), 303–308.

Birolini, A. (2012). *Quality and reliability of technical systems: Theory, practice, management.* Springer Science & Business Media.

Boudreau, M. C., Gefen, D., & Straub, D. W. (2001). Validation in information systems research: A state-of-the-art assessment. *MIS Quarterly, 25*(1), 1–16.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*(3), 687–699.

Breward, M., Hassanein, K., & Head, M. (2017). Understanding consumers' attitudes toward controversial information technologies: A contextualization approach. *Information Systems Research, 28*(4), 760–774.

Brown, R. D., & Hauenstein, N. M. (2005). Interrater agreement reconsidered: An alternative to the rwg indices. *Organizational Research Methods, 8*(2), 165–184.

Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., & Abel, T. (2008). From text to codings: Intercoder reliability assessment in qualitative content analysis. *Nursing Research, 57*(2), 113–117.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*(5), 423–429.

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

Cheng, X., Fu, S., & de Vreede, G. J. (2017). Understanding trust influencing factors in social media communication: A qualitative study. *International Journal of Information Management, 37*(2), 25–35.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Compton, D. L., Love, T. P., & Sell, J. (2012). Developing and assessing intercoder reliability in studies of group interaction. *Sociological Methodology, 42*(1), 348–364.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin, 88*(2), 322–328.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

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.

Davies, K. (2012). Content analysis of research articles in information systems (LIS) journals. *Library and Information Research, 36*(112), 16–28.

De Swert, K. (2012). Calculating inter-coder reliability in media content analysis using Krippendorff's alpha. *Center for Politics and Communication,* 1–15.

Dong, M. C., Fang, Y., & Straub, D. W. (2017). The impact of institutional distance on the joint performance of collaborating firms: The role of adaptive interorganizational systems. *Information Systems Research, 28*(2), 309–331.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., et al. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management,* 85–95.

Feng, G. C. (2014a). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology, 11*, 13–22.

Feng, G. C. (2014b). Intercoder reliability indices: Disuse, misuse, and abuse. *Quality & Quantity, 48*(3), 1803–1815.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382.

Gaskin, J. E., Berente, N., Lyytinen, K., & Yoo, Y. (2014). Toward generalizable sociomaterial inquiry: A computational approach for zooming in and out of sociomaterial

routines. *MIS Quarterly, 38*(3), 849–871.
Gerdes, J., Stringam, B. B., & Brookshire, R. G. (2008). An integrative approach to assess qualitative and quantitative consumer feedback. *Electronic Commerce Research, 8*(4), 217–234.
Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly, 37*(2), 337–355.
Grover, V., & Lyytinen, K. (2015). New state of play in information systems research: The push to the edges. *MIS Quarterly, 39*(2), 271–296.
Gwet, K. (2002). Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series, 2*, 1–9.
Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *The British Journal of Mathematical and Statistical Psychology, 61*, 29–48.
Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* Advanced Analytics, LLC.
Hallgren, K. A. (2014). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23–34.
Hayes, A. F. (2005). *Statistical methods for communication science.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Hayes, A. F. (2009). *Statistical methods for communication science.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89.
Holley, W., & Guilford, J. P. (1964). A note on the G-Index of agreement. *Educational and Psychological Measurement, 24*, 749–753.
Holsti, O. R. (1969). *Content analysis for the social sciences and humanities.* Reading, MA: Addison-Wesley.
Hruschka, D. J., Schwartz, D., St. John, D. C., Picone-Decaro, E., Jenkins, R. A., & Carey, J. W. (2004). Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field Methods, 16*(3), 307–331.
Hsieh, J. K., & Hsieh, Y. C. (2013). Appealing to Internet-based freelance developers in smartphone application marketplaces. *International Journal of Information Management, 33*(2), 308–317.
Hughes, M. A., & Garrett, D. E. (1990). Intercoder reliability estimation approaches in marketing: A generalizability theory framework for quantitative data. *Journal of Marketing Research, 27*(2), 185–195.
International Journal of Information Management (2019). https://www.journals.elsevier.com/international-journal-of-information-management Accessed 25 May 2019.
Ismagilova, E., Hughes, L., Dwivedi, Y. K., & Raman, K. R. (2019). Smart cities: Advances in research—An information systems perspective. *International Journal of Information Management, 47*, 88–100.
James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *The Journal of Applied Psychology, 69*(1), 85–98.
Janson, S., & Vegelius, J. (1979). On generalizations of the G Index and the Phi coefficient to nominal scales. *Multivariate Behavioral Research, 14*(2), 255–269.
Karhade, P., Shaw, M. J., & Subramanyam, R. (2015). Patterns in information systems portfolio prioritization: Evidence from decision tree induction. *MIS Quarterly, 39*(2), 413–433.
King, N. T. G. S., & Cassell, C. (Eds.). (1998). *Qualitative methods and analysis in organizational research: A practical guide* (pp. 118–134). Thousand Oaks, CA: Sage Publications.
Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30*(1), 61–70.
Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.
Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures, 5*(2), 93–112.
Krippendorff, K. (2013). Commentary: A dissenting view on so-called paradoxes of reliability coefficients. *Annals of the International Communication Association, 36*(1), 481–499.
Krippendorff, K. (2012). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, Calif: Sage.
Kurasaki, K. S. (2000). Intercoder reliability for validating conclusions drawn from open-ended interview data. *Field Methods, 12*(3), 179–194.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 59–174.
Li, L., Zhang, Q., Tian, J., & Wang, H. (2018). Characterizing information propagation patterns in emergencies: A case study with Yiliang Earthquake. *International Journal of Information Management, 38*(1), 34–41.
Lindberg, A., Berente, N., Gaskin, J., & Lyytinen, K. (2016). Coordinating interdependencies in online communities: A study of an open source software project. *Information Systems Research, 27*(4), 751–772.
Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*(4), 587–604.
Ludwig, S., De Ruyter, K., Mahr, D., Wetzels, M., Brüggen, E., & De Ruyck, T. (2014). Take their word for it: The symbolic role of linguistic style matches in user communities. *MIS Quarterly, 38*(4), 1201–1217.
MacPhail, C., Khoza, N., Abler, L., & Ranganathan, M. (2016). Process guidelines for establishing intercoder reliability in qualitative studies. *Qualitative Research, 16*(2), 198–212.
Mäntymäki, M., & Salo, J. (2015). Why do teens spend real money in virtual worlds? A consumption values and developmental psychology perspective on virtual consumption. *International Journal of Information Management, 35*(1), 124–134.
Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *The British Journal of Psychiatry, 130*(1), 79–83.
Miles, M., Huberman, A., & Saldana, J. (2014). *Qualitative data analysis: A methods sourcebook.* Thousand Oaks, CA: Sage.
MiniTab Inc (2019). *Kappa statistics for attribute agreement analysis.* Accessed 22 Jan 2019 https://support.minitab.Inc.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/how-to/attribute-agreement-analysis/attribute-agreement-analysis/interpret-the-results/all-statistics-and-graphs/kappa-statistics/.
Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2002). Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods, 1*(2), 13–22.
Nili, A., Tate, M., & Johnstone, D. (2017). A framework and approach for analysis of focus group data in information systems research. *Communications of the Association for Information Systems, 40*(1), 1–21.
Oh, O., Eom, C., & Rao, H. R. (2015). Research note—Role of social media in social change: An analysis of collective sense making during the 2011 Egypt revolution. *Information Systems Research, 26*(1), 210–223.
Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly, 37*(2), 407–426.
Olson, J. D., McAllister, C., Grinnell, L. D., Walters, K. G., & Appunn, F. (2016). Applying constant comparative method with multiple investigators and inter-coder reliability. *The Qualitative Report, 21*(1), 26–42.
Ou, C. X., Pavlou, P., & Davison, R. (2013). Swift guanxi in online marketplaces: The role of computer-mediated communication technologies. *MIS Quarterly, 38*(1), 209–230.
Palese, B., & Usai, A. (2018). The relative importance of service quality dimensions in e-commerce experiences. *International Journal of Information Management, 40*, 132–140.
Park, S., & Park, K. (2015). Intercoder reliability indices in tourism research. *Annals of Tourism Research, 55*, 180–183.
Pearson, K., Lee, A., Warren, E., Fry, A., & Fawcett, C. D. (1901). Mathematical contributions to the theory of evolution. IX.–on the principle of homotyposis and its relation to heredity, to the variability of the individual, and to that of the race. Part I.–Homotyposis in the vegetable kingdom. *Proceedings of the Royal Society of London, 68*, 442–450 1-5.
Perreault, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research, 26*(2), 135–148.
Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27*(3), 258–284.
Rogot, E., & Goldberg, I. D. (1966). A proposed index for measuring agreement in test-retest studies. *Journal of Chronic Diseases, 19*(9), 991–1006.
Ruggeri, A., Gizelis, T. I., & Dorussen, H. (2011). Events data as Bismarck's sausages? Intercoder reliability, coders' selection, and data quality. *International Interactions, 37*(3), 340–361.
Rust, R., & Cooil, B. (1994). Reliability measures for qualitative data: Theory and implications. *Journal of Marketing Research, 31*(1), 1–14.
Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*(3), 321–325.
Siegel, S., & Castellan, N. J. (1988). *The case of k related samples. Nonparametric Statistics for Behavioral Sciences.* New York: McGraw-Hill170–174.
Stevens, M. R., Lyles, W., & Berke, P. R. (2014). Measuring and reporting intercoder reliability in plan quality evaluation research. *Journal of Planning Education and Research, 34*(1), 77–93.
Tamilmani, K., Rana, N. P., Prakasam, N., & Dwivedi, Y. K. (2019). The battle of brain vs. heart: A literature review and meta-analysis of "hedonic motivation" use in UTAUT2. *International Journal of Information Management, 46*, 222–235.
Tan, C. W., Benbasat, I., & Cenfetelli, R. T. (2016). An exploratory study of the formation and impact of electronic service failures. *MIS Quarterly, 40*(1), 1–29.
The Belmont Report (1978). *Ethical principles and guidelines for the protection of human subjects of research. The national commission for the protection of human subjects of biomedical and behavioral research. DHEW publication No. (OS) 78-0012.* Accessed 30 March 2020 https://repository.library.georgetown.edu/bitstream/handle/10822/779133/ohrp_belmont_report.pdf?sequence=1&isAllowed=y.
Tsai, H. T., & Bagozzi, R. P. (2014). Contribution behavior in virtual communities: Cognitive, emotional, and social influences. *MIS Quarterly, 38*(1), 143–163.
Venable, J., & Baskerville, R. (2012). Eating our own cooking: Toward a design science of research methods. *Proceedings of the 11th European Conference on Research Methods* (pp. 399–407).
Venkatesh, V., Brown, S. A., & Bala, H. (2013). Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems. *MIS Quarterly, 37*(1), 21–54.
Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology, 16*(1), 93.
Zhang, C., Fan, C., Yao, W., Hu, X., & Mostafavi, A. (2019). Social media for intelligent public information and warning in disasters: An interdisciplinary review. *International Journal of Information Management, 49*, 190–207.
Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association, 36*(1), 419–480.