2017

A Critical Analysis of Inter-Coder Reliability Methods in Information Systems Research

Alireza Nili
Queensland University of Technology, a.nili@qut.edu.au

Mary Tate
Queensland University of Technology, mary.tate@qut.edu.au

Alistair Barros
Queensland University of Technology, alistair.barros@qut.edu.au

A Critical Analysis of Inter-Coder Reliability Methods in Information Systems Research

Alireza Nili
School of Information Systems
Queensland University of Technology
Brisbane, Australia
Email: a.nili@qut.edu.au

Mary Tate
School of Information Systems
Queensland University of Technology
Brisbane, Australia
Email: mary.tate@qut.edu.au

Alistair Barros
School of Information Systems
Queensland University of Technology
Brisbane, Australia
Email: alistair.barros@qut.edu.au

Abstract

Qualitative researchers in business and management information systems fields often need to employ a method of inter-coder reliability to test the trustworthiness of the findings of their content analysis. A suitable method for checking the inter-coder reliability enables researchers to rigorously assess the degree of agreement among two or more independent qualitative coders. By employing this method, researchers can 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. However, little guidance is available on what method of inter-coder reliability check should be used. In this paper, we present a critical analysis of these methods that are suitable for qualitative business and management IS research, and provide an example of how we employed the most rigorous method among these methods for a qualitative behavioural IS study.

Keywords: Inter-coder reliability, content analysis, behavioural information systems, business and MIS

1 Introduction

Qualitative coding can be conducted for many purposes and in the context of different research paradigms. More constructivist and relativist research paradigms may focus more on the authenticity with which the individual researcher allows the data to speak, and less on the degree of agreement among multiple qualitative coders (Lincoln and Guba 1986). This is because the researcher themselves, and their own value system, are recognised as an integral part of the research process. However, many qualitative coding exercises are conducted within a post-positivist paradigm (Dube and Pare 2003). Even when the enquiry is initially inductive, sometimes later stages in the data analysis seek in some way to confirm the emerging theoretical framework derived from the data, for example in the selective coding steps of a grounded study (Strauss and Corbin 1990). Overall, in any situation where qualitative data analysis is presented as evidence for the salience of a theoretical framework which claims some ontological existence that is separate from the researcher's interpretation and construction, it is common to seek some evidence for the reliability of the coding effort. This is frequently achieved by comparing the results of coding carried out independently by two or more coders (Sanders and Cuneo 2010; Nili et al. 2017).

A broad range of information systems (IS) studies employ a method of collecting qualitative data, such as semi-structured interviews and focus groups, and a content analysis of this data is an important part of these studies (Boudreau et al. 2001; Davies 2012). In content analysis, data in the transcripts (audio and/or video records of participant's responses converted into text) are typically coded by trained qualitative coders, and these codes can be trusted only after ensuring their reliability (Davies 2012; Nili et al. 2017; Venkatesh et al. 2013). In order to ensure reliability of the coding, a method of checking inter-coder reliability of data needs to be employed. Such a method enables researchers to assess the degree of agreement among two or more independent coders on the accuracy and correct allocation of the codes to the right data chunks (relevant participant's comments, response or opinion) in the transcript. The more coders agree on the codes, the more confident we can be that the codes are exchangeable with codes provided by another coder (i.e. the findings are reproducible and trustworthy) (Davies 2012; Olson et al. 2016).

By employing a suitable method of inter-coder reliability check, 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 avoid any associated major time, effort and financial cost (Morse et al. 2002; Venkatesh et al. 2013; Nili 2016). This is particularly important for IS studies which require a high level of precision in content analysis (e.g. in many studies in the area of health IS where a high level of sensitivity, risk and ethics is involved). "High reliability makes it less likely that bad managerial decisions will result from using the data" (Rust and Cooil 1994 p.11).

"Researchers who conduct qualitative research have long faced the challenge of providing appropriate reliability" (Park and Park 2015 p.180). Choosing a method of inter-coder reliability check is complicated as there are a number of possible approaches, and there is no consensus among researchers on which method is the best one (Davies 2012; Venkatesh et al. 2013; Olson et al. 2016). Also, there is considerable confusion between the terms "inter-rater reliability" and "inter-coder reliability". Inter-rater reliability refers to a situation where the codes and the values that can be assigned to them are already defined. Two or several scorers or judges independently award a value, and their results are compared (Gwet 2014; Hallgren 2014). Inter-coder reliability refers to the situation where coders are requested to relate pre-defined codes to related data chunks in a transcript to independently validate coding, and the level of agreement among coders is measured (Campbell et al. 2013; MacPhail et al. 2016). There is very little consistency in the use of these terms, and some texts refer to inter-coder reliability when what they describe most closely resembles inter-rater reliability. Although checking the inter-coder reliability is a significant component of content analysis in many IS studies, choosing the right method is still a significant challenge for IS researchers. In this study, we elucidate *inter-coder* reliability. Readers should be aware of the difference between the two, and the widespread confusion in terminology, which can result in a misapplication of methods.

In this paper, we present a critical analysis of the methods of inter-coder reliability checking that are suitable for use in IS research; and provide an example of how we selected and employed a method for a qualitative IS study. To do this, we first present a criteria for assessing the main characteristics of the methods of inter-coder reliability check based on our literature review. Next, we present an analysis of the methods based on this criteria. We then present a method for selecting and applying an inter-coder reliability method for the content analysis for a behavioural IS study, supported by a running example. The paper ends with the discussion and conclusion section.

2 Criteria for a Good Method of Inter-Coder Reliability Check

There are more than thirty different statistical methods of inter-coder reliability checking, but only a few of them have been widely accepted and used, and there is no consensus among researchers on the single best one (Feng 2014a; 2014b; Olson et al. 2016; Sanders and Cuneo 2010), or even clear guidance as to the appropriate selection of a method depending on the requirements of the research. We reviewed conference and journal papers (regardless of their rank) that have been written specifically about checking the inter-coder reliability and the research methodology books in IS, business and fields related to the broad social sciences research (e.g. communication and media studies) that we have access from the digital library of our university in Australia. We only considered the studies which have focused on providing a detailed description of these methods or have provided a detailed criticism of a method. We identified Feng (2014a, 2014b), Hayes and Krippendorff (2007), Lombard et al. (2002), Olson et al. (2016), Sanders and Cuneo (2010), Stevens et al. (2014) and Zhao et al. (2013) (none of this studies is in the IS field), as the most relevant papers to the aim of this study.

According to the selected papers, particularly according to Feng (2014a; 2014b), Hayes and Krippendorff (2007), Krippendorff (2013) and Sanders and Cuneo (2010), who have covered using more than one method of inter-coder reliability in their studies, researchers need to consider the criteria below when they want to select a method to check inter-coder reliability of their content analysis:

i. The type(s) of data (nominal, ordinal, interval and ratio) the method is applicable for.

ii. Whether the method allows assessing agreement between two independent or among more than two independent coders. Content analysis of an IS study in which a high level of risk and sensitivity is involved in its findings and requires a high level of precision may require more than two independent coders. Examples of these include many studies in the area of health IS where a high level of sensitivity, risk and human ethics is involved and a project management study in which significant financial cost will be incurred if a wrong decision is made based on the findings.

iii. Whether the method calculates or minimises the role of 'chance' in the independent coders' agreement on a code. Similarly, this is particularly important for an IS study, in which the content analysis requires a high level of precision in its content analysis.

iv. Where the result should be reported quantitatively, it can be represented by a numerical scale, such as 1.000 or 100% showing the perfect agreement, and 0.000 showing complete disagreement among coders.

v. Detailed information on how the method is used can be communicated clearly (the method allows for providing a transparent procedure of how it was used).

Also, it is important to note that in terms of procedure of using the method, independent coders should have the latitude to identify any new codes in the transcripts (Campbell et al. 2013; Krippendorff 2013). Below, we present a brief analysis of these methods. In this analysis, we do not present mathematical details of the methods. Instead, we refer interested readers to the original sources for these methods.

3 An Analysis of the Methods of Inter-Coder Reliability Check

In this section we present a critical review of the inter-coder reliability methods.

Percent Agreement: *Percent agreement* is the most widely used method in various different research fields, mainly because of its ease of use and simplicity (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, although it is simple to transparently communicate how the method is used in a study, the method is perceived as an unreliable method for checking agreement among coders in a study that requires a high level of precision in analysis (Krippendorff 2013; Zhao et al. 2013). Percent agreement and Holsti's CR, which is a variation of the percent agreement method, may overestimate true agreement between coders (Feng 2014a; 2014b; Zhao et al. 2013). As the number of codes increase, however, high agreement between coders becomes more difficult, allowing for argument on the suitability of percent agreement for such a 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 type of data. Compared with percent agreement, however, the method is considered 'more'

reliable (but not at an optimum level) in terms of considering chance in agreement (Hayes and Krippendorff 2007; Lombard et al. 2002). It can be used regardless of complexity of the content analysis (e.g. the number of codes does not affect calculation of reliability 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 and Krippendorff 2007 p.80). In the past three decades, the method has been revised several times as the inter-coder reliability coefficient *Ir*. (Perreault and 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 (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.

Fleiss's *K:* Fleiss (1971) generalised 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 agreement 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). Compared with percent agreement, the method is a more 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 and Krippendorff 2007; Olson et al. 2016).

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. What is considered as a high, acceptable or low agreement is unclear. The only clear aspect of its outcome is that 1 implies perfect agreement and any result less than 1 implies a not perfect agreement (Olson et al. 2016; Zhao et al. 2013). 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's (1966) A_2 coefficient is equivalent to the Cohen's kappa method.

Gwet (2014): In a more recent work, Gwet (2014), revised and generalised Bennett et al.'s S, Scott's pi, Fleiss's *K* and Cohen's Kappa to accommodate multiple coders and missing codes that my occur as the result of mistake in coding of data by a coder[1]. 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 and Guilford 1964); Brennan and Prediger's (1989) free marginal kappa coefficient; Byrt et al.'s (1993) prevalence and bias adjusted kappa coefficient; Jason and Vegelius's (1979) C score; Maxwell's (1977) random error coefficient; and Potter and Levine-Donnerstein's (1999) redefined pi coefficient.

Krippendorff's alpha: Krippendorff's alpha (1970; 2004a) resolves the limitations of the other methods of inter-coder reliability check. 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 minimises the effect of chance in agreements on the codes (Feng 2014a; 2014b; Park and Park 2015). These characteristics of the method are mainly because the method calculates disagreements among coders instead of correcting percent agreements, and because it embraces several known reliability coefficients, including Scott's π for its two-coder nominal data calculation; Pearson et al.'s (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 (Lombard et al., 2002; Feng 2014a; 2014b). Krippendorff

[1] We note that a few previous versions of Gwet's (2002; 2008; 2010) method (sometimes called AC1 and AC2) also exist, but they have not been adopted significantly by IS, business, and any field in the social sciences (Feng 2014b).

(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.

Cronbach's alpha, Chi-square and Pearson's r: We note that chi-square (which measures association), Cronbach's alpha (or αC which measures internal consistency; Cronbach 1951), and Pearson's r (which measures correlation) should not be confused with methods of inter-coder reliability check, as none of these three types of methods measure 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.

4 Comparing Methods of Inter-Coder Reliability Check

Choosing a method of inter-coder reliability check for a study needs assessing the method based on its properties and the characteristics of the content analysis of the study. For example, consider a study which needs only two coders to code only the nominal type of data and its content analysis does not require a high level of precision (e.g. where no risk to health or a significant financial cost will be incurred if a wrong decision is made based on the findings). Such a 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.

Table 1 presents a summary of the characteristics of these methods. As the table shows, percent agreement, which is the most widely used method of inter-coder reliability check, is the least flexible method. It is applicable only for nominal data, it does not consider missing codes, and there is no general agreement on what percentage is a suitable enough level of agreement. It is not recommended as a suitable method for a study which requires a high degree of precision in its analysis, since it does not consider the effect of chance in agreement and does not allow for more than two coders to be involved in the content analysis. However, even the percent agreement method and its extended version (Holsti's CR) could be appropriate enough if the content analysis of the study does not require a higher degree of interpretation of data (i.e. the easier the content analysis is, the less is the chance of agreement on the codes by independent coders) and where there are many codes (i.e. the more codes the coders need to agree on, the less probable is the coders' agreement on the codes that could occur by chance). Also, as the table shows, Krippendorff's alpha is very flexible and produces as maximally accurate results. In other words, it can be used with more than two coders for ordinal, nominal, interval and ratio data and even missing data, and is very suitable for the studies that require a high level of precision in their content analysis. Whichever method is employed, researchers should briefly explain why the characteristics of the method are appropriate for the specific characteristics of their study.

Method	Type of data	Missing codes?	Number of coders	The effect of 'chance' in agreement is minimised?	General agreement on the significance of a numeric result?*
Percent Agreement	Nominal	No	Two	No	No
Bennett et al.'s S	Nominal	No	Two	No**	No
Scott's Pi	Nominal	No	Two***	No**	No
Fleiss's *K*	Nominal	No	Multiple	No**	No
Cohen's Kappa	Nominal	No	Two****	No**	No
Gwet (2014)	Nominal	Yes	Multiple	No**	No (mainly due to its newness)
Krippendorff's Alpha	All four types	Yes	Multiple	Yes	Yes

* e.g. if Krippendorff's alpha is used, 0.823 (which is between 0.8 and 0.9) means a suitable amount of inter-coder reliability.
** with regards to reducing the effect of chance in agreement, the method has a better status compared to percent agreement and to Holsti's CR – the revised version of percent agreement.
*** the extended version by Siegel and Castellan (1988) accommodates multiple coders.
**** the extended version by Conger (1980) accommodates multiple coders.

Table 1. A Comparison of the Characteristics of Inter-Coder Reliability Methods

We note that before using any of the methods, researchers may need to have a small training session with the coders, where the researcher explains the purpose and background of the research, explains the coding schemes and clarifies the definition of each code, and conducts a couple of practice runs with a small sample of data. Such a practice and training can particularly be beneficial where the content analysis requires a high level of interpretation of data.

4.1 Use of Inter-Coder Reliability Method in Information Systems Research

We reviewed the qualitative papers and papers which have used a mixed-method approach from the four most recent issues of MIS Quarterly, and identified seven papers which include a content analysis of data. Among these, only three papers have discussed the selection of application of inter-coder reliability methods. These include Benaroch and Chernobai (2017) who used percent agreement, and Chen et al. (2017) and Jensen and Yetgin (2017) who used Cohen's Kappa. Each of these papers have employed two coders for their research. We do not presume to comment in the overall quality of these seven research papers. As we noted earlier, issues of research design and paradigmatic assumptions may affect the degree to which these considerations are important to the study. However, the absence of any discussion of this issue in recent papers in our leading journal is a further motivation for the importance of our study.

5 An Example of Selecting and Using an Inter-coder Reliability method

In this section, we present an illustrative example for selecting and applying an inter-coder reliability method for the content analysis of a qualitative IS study.

5.1 Develop Coding Scheme

The first step is to develop a coding scheme. This may be done a priori, or inductively and iteratively from data. However, eventually, the coding scheme will be finalised and the definition of each code will be established.

Example: The project aimed to identify the factors that contribute to user persistence in solving IT problems. An a priori coding scheme was developed based on an extensive literature review. The outcome of the project was intended to be a generalisable theoretical model in a post-positivist paradigm. Codes were intended to be able to be 1) compared with factors identified in literature; and 2) used to guide the development of quantitative constructs. Overall, 30 individual interviews were conducted with users who had experienced IT problems. These were then transcribed carefully by the researchers. As the result of initial content analysis, the research team identified 33 codes. These were largely based on the a priori framework, although some modifications were made. By a code we mean a 'label' for each factor that contributes to user persistence. The label is in the form of one or a few words (e.g. "relevancy" of self-help information and "ease of use" of technology) that describes data chunks about the contributing role of the factor to user persistence (Nili et al. 2014; 2017).

5.2 Select and Train Independent Coders

The next step, once the coding scheme and definitions are clearly established and agreed by the research team and when initial coding has been completed, is to select and train the independent coders. Their work will be used to calculate inter-coder reliability.

Example: A table was developed in which each code (or label) had a number and a specific definition. Two independent coders with previous qualitative coding experience but no previous experience with the project were selected. The original research team explained the purpose of the research and presented the coding sheet. At the meeting, coding practice was carried out with a small sample of the data, where we asked each coder to independently allocate the number of each code to its related data chunk in the transcripts. When the coders were feeling confident about this coding process, coders were asked to code half of the data independently. An initial screening of the two coder's coding sheet showed a high level of agreement between the coders and showed that coders were feeling confident about their coding skills, making the team confident that further training is not required. Therefore, the coders were asked to code the rest of the transcript.

5.3 Select Method for Evaluating Inter-Coder Reliability

In this step, the nature of the data and coding scheme, number of coders, and level of precision in content analysis are considered, and the appropriate method is selected.

Example: Because of the type of data, the number of independent coders (two coders), not having any missing codes, and not considering a high level of precision in our content analysis, we could have selected any of the seven methods for our research. However, we chose Krippendorff's alpha, since among all methods of checking the inter-coder reliability, Krippendorff's alpha is a very flexible method that produces the most possible accurate result.

5.4 Calculate Inter-Coder Reliability

In this step, the selected method to calculate inter-coder reliability is used, based on the results of the coding. We recommend to use a software that is able to support the calculation.

Example: Having selected Krippendorff's alpha as the most appropriate method, in order to calculate it, we entered a number for each data chunk into SPSS software (see the left column in Table 2 which present a portion of data on the 'data view' page of the software). Next, the code number (e.g. 3 or 16) that each coder allocated to each data chunk was entered in the second and the third column for that data chunk (as mentioned earlier, we had allocated a number for each code in the coding scheme and asked coders to allocate 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. Hayes and Krippendorff 2007; De Swert 2012) which have used Krippendorff's alpha to check inter-*rater* reliability.

DataChunk	Coder1	Coder2
1	3	3
2	2	2
3	6	6
4	12	12
5	16	16
6	21	21
7	8	8
8	7	7
9	9	9
10	10	10
11	21	21
12	20	20

Table 2. A Portion of 'Data View' Page of SPSS

Although some software and applications such as PRAM allows calculation of Krippendorff's alpha easily, we chose SPSS, as the software is very popular among behavioural IS researchers. The software however does not do this calculation directly, and requires installing a macro which was developed by Hayes (2005; 2009) to make this calculation possible[2].

5.5 Interpret the Result

In this step, the results of the inter-coder reliability check are interpreted based on most recent heuristics and standards.

Example: The result of the inter-coder reliability check by the software was 0.877 which is considered suitable, especially for a project with 33 codes which is a significant number. The project used two fully independent qualitative coders. The way the method is used is however the same when multiple coders are employed. Overall, the research team was confident in the reliability of the results of the coding exercise.

6 Discussion and Conclusion

Choosing an appropriate method of inter-coder reliability check for a study needs assessing the method based on its properties and the characteristics of the content analysis of the study. We recommend that whichever method is employed, researchers need to explain why the characteristics of the method are appropriate for the specific characteristics of their study. Considering a range of study characteristics,

[2] Interested researchers can download the macro by looking for KALPHA.sps from http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html

on balance, Krippendorff's alpha should be considered as the method of choice, as it is very flexible and produces more accurate results. This is particularly important for studies which require a high level of precision and reliability in their content analysis. However, compared to other methods, Krippendorff's alpha is usually perceived as a tedious method and the lack of implementation of the method in popular statistical packages has negatively affected the full utilisation of its capabilities. Therefore, in our practical example, we used SPSS and provided a reference to a macro that is written for the software and for SAS to compute Krippendorff's alpha.

In our evaluation of a small sample of recent papers from a leading IS journal to obtain a 'snapshot' of current practices with regards to evaluating inter-coder reliability, we identified only three papers that provided a discussion of the method they used for calculating inter-coder reliability. The absence of any discussion of this issue in recent papers in a leading IS journal is a further motivation for the importance of our study. Future research could carry out a more comprehensive evaluation of IS papers in leading journals to examine how authors have approached the selection and reporting of methods for calculating inter-coder reliability, including recommendations for improvement. For example, this assessment can consider whether the papers have justified their choice of the method transparently, whether they have mentioned the number of coders, whether they have used a method for the right type of data, and whether they needed to and considered any missing codes in checking their inter-coder reliability.

Considering the widespread use of qualitative coding techniques using a priori codes (or codes developed at an early stage in the research process) to support, or test theoretical frameworks, establishing the reliability of these codes is essential to establishing the credibility and trustworthiness of the research. It is therefore surprising (and concerning) that this issue has not received more attention, as inter-coder reliability can be considered a key quality metric for many IS studies which use a content analysis of qualitative data.

We believe that employing a suitable method of inter-coder reliability check by both experienced and novice researchers will definitely pay off the initial cost. By employing a suitable method of inter-coder reliability check, researchers can identify and correct mistakes in their content analysis before the codes are used in developing and testing a theory, theoretical framework, or a measurement model.

Overall, the contribution of this paper is to: 1) clarify the nature and importance of evaluating inter-coder reliability in the qualitative research process; 2) review leading methods of inter-coder reliability check that are suitable for use in IS research; 3) present guidance on selecting a suitable method of inter-coder reliability check; 4) recommend Krippendorff's alpha as the method of choice; and 5) provide an approach for carrying out inter-coder reliability on a qualitative data analysis project, illustrated with a detailed example providing an example of how the method can be used in a qualitative IS study.

References

Benaroch, M., and Chernobai, A. 2017. "Operational IT Failures, IT Value-Destruction, and Board-Level IT Governance Changes," *MIS Quarterly*, forthcoming.

Bennett, E. M., Alpert, R., and Goldstein, A. C. 1954. "Communications through Limited-Response Questioning. *Public Opinion Quarterly*," (18:3), pp 303-308.

Boudreau, M. C., Gefen, D., and Straub, D. W. 2001. "Validation in Information Systems Research: A State-of-the-Art Assessment," (25:1), *MIS Quarterly*, pp 1-16.

Brennan, R. L., and Prediger, D. J. 1981. "Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement,*" (41:3), pp 687–699.

Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., and Abel, T. 2008. "From Text to Codings: Intercoder Reliability Assessment in Qualitative Content Analysis," *Nursing Research* (57:2), pp 113-117.

Byrt, T., Bishop, J., and Carlin, J. B. 1993. "Bias, Prevalence and Kappa," *Journal of Clinical Epidemiology* (46:5), pp 423-429.

Campbell, J. L., Quincy, C., Osserman, J., and Pedersen, O. K. 2013. "Coding In-Depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement," *Sociological Methods & Research* (42:3), pp 294-320.

Chen, Y., Bharadwaj, A., and Goh, K. Y. 2017. "An Empirical Analysis of Intellectual Property Rights Sharing in Software Development Outsourcing," *MIS Quarterly* (41:1), pp 131-161.

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

Compton, D. L., Love, T. P., and Sell, J. 2012. "Developing and Assessing Intercoder Reliability in Studies of Group Interaction," *Sociological Methodology* (42:1), pp 348-364.

Conger, A. J. 1980. "Integration and Generalization of Kappas for Multiple Raters," *Psychological Bulletin* (88:2), pp 322-328.

Cronbach, L. J. 1951. "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika* (16:3), pp 297-334.

Davies, K. 2012. "Content Analysis of Research Articles in Information Systems (LIS) Journals," *Library and Information Research* (36:112), pp 16-28.

De Swert, K. 2012. "Calculating Inter-Coder Reliability in Media Content Analysis Using Krippendorff's Alpha," *Center for Politics and Communication*, pp 1-15.

Dube, L., and Pare, G. 2003. "Rigor in Information Systems Positivist Case Research: Current Practices, Trends, and Recommendations," *MIS Quarterly* (27:4), pp 597-635.

Feng, G. C. 2013. "Factors Affecting Intercoder Reliability: A Monte Carlo experiment," *Quality & Quantity* (47:5), pp 1-24. DOI 10.1007/s11135-012-9745-9

Feng, G. C. 2014a. "Mistakes and How to Avoid Mistakes in Using Intercoder Reliability Indices," *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*. Advance online publication. http://dx.doi.org/10.1027/1614-2241/a000086

Feng, G. C. 2014b. "Intercoder Reliability Indices: Disuse, Misuse, and Abuse," *Quality & Quantity* (48:3), pp 1803-1815.

Fleiss, J. L. 1971. "Measuring Nominal Scale Agreement among Many Raters," *Psychological Bulletin* (76:5), pp 378-382.

Gwet, K. 2002. "Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity," *Statistical Methods for Inter-Rater Reliability Assessment Series* (2), pp 1-9.

Gwet, K. L. 2008. "Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement," *British Journal of Mathematical and Statistical Psychology* (61), pp 29-48.

Gwet, K. L. 2010. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Multiple Raters* (2nd ed.). Gaithersburg, MD: Advanced Analytics.

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," *Tutor Quant Methods Psychol* (8:1), pp 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., and Krippendorff, K. 2007. "Answering the Call for a Standard Reliability Measure for Coding Data," *Communication Methods and Measures* (1:1), pp 77-89.

Holley, W., and Guilford, J. P. 1964. "A Note on the *G*-Index of Agreement," *Educational and Psychological Measurement* (24), pp 749–753.

Hruschka, D. J., Schwartz, D., St. John, D. C., Picone-Decaro, E., Jenkins, R. A., and Carey, J. W. 2004. "Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research," *Field Methods* (16:3), pp 307-331.

Janson, S., and Vegelius, J. 1979. "On Generalizations of the G Index and the Phi Coefficient to Nominal Scales," *Multivariate Behavioral Research* (14:2), pp 255-269.

Jensen, M. L., and Yetgin, E. 2017. "Prominence and Interpretation of Online Conflict of Interest Disclosures," *MIS Quarterly* (41:2).

Krippendorff, K. 1970. "Estimating the Reliability, Systematic Error and Random Error of Interval Data," *Educational and Psychological Measurement* (30:1), pp 61-70.

Krippendorff, K. 2004a. *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), pp 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), pp 179-194.

Lincoln, Y., and Guba, E. 1986. *But is it Rigorous? Trustworthiness and Authenticity in Naturalistic Evaluation*. In D. Williams (Ed.), *New Directions for Program Evaluation* (30). SanFrancisco: Jossey-Bass.

Lombard, M., Snyder-Duch, J., and Bracken, C. C. 2002. "Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability," *Human Communication Research* (28:4), pp 587-604.

MacPhail, C., Khoza, N., Abler, L., and Ranganathan, M. 2016. "Process Guidelines for Establishing Intercoder Reliability in Qualitative Studies," *Qualitative Research* (16:2), pp 198-212.

Maxwell, A. E. 1977. "Coefficients of Agreement between Observers and Their Interpretation," *British Journal of Psychiatry* (130:1), pp 79-83.

Morse, J. M., Barrett, M., Mayan, M., Olson, K., and Spiers, J. 2002. "Verification Strategies for Establishing Reliability and Validity in Qualitative Research," *International Journal of Qualitative Methods* (1:2), pp 13-22.

Nili, A. 2016. *Digital Service Problems: Prevention and User Persistence in Solving Them*. Doctoral thesis submitted to Victoria University of Wellington.

Nili, A., Tate, M., Johnstone, D., and Gable, G.G. 2014. "A Framework for Qualitative Analysis of Focus Group Data in Information Systems," In *25th Australasian Conference on Information Systems*, pp 8-10 December 2014, Auckland, New Zealand.

Nili, A., Tate, M., and 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), pp 1-21.

Olson, J. D., McAllister, C., Grinnell, L. D., Walters, K. G., and Appunn, F. 2016. "Applying Constant Comparative Method with Multiple Investigators and Inter-Coder Reliability," *The Qualitative Report* (21:1), pp 26-42.

Park, S., and Park, K. 2015. "Intercoder Reliability Indices in Tourism Research," *Annals of Tourism Research* (55), pp 180-183.

Pearson, K., Lee, A., Warren, E., Fry, A., and 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(pp 442-450), 1-5.

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

Potter, W. J., and Levine-Donnerstein, D. 1999. "Rethinking Validity and Reliability in Content Analysis," *Journal of Applied Communication Research* (27:3), pp 258-284.

Riff, D., Lacy, S., and Fico, F. G. 2005. *Analyzing Media Messages: Using Quantitative Content Analysis in Research* (2nd ed.). Mahwah, NJ: Erlbaum.

Rogot, E., and Goldberg, I. D. 1966. "A Proposed Index for Measuring Agreement in Test-Retest Studies," *Journal of Chronic Diseases* (19:9), pp 991-1006.

Rubin, V. L. 2007. Stating with Certainty or Stating with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers* (pp 141-144). Association for Computational Linguistics.

Ruggeri, A., Gizelis, T. I., and Dorussen, H. 2011. "Events Data as Bismarck's Sausages? Intercoder Reliability, Coders' Selection, and Data Quality," *International Interactions* (37:3), pp 340-361.

Rust, R., and Cooil, B. 1994. "Reliability Measures for Qualitative Data: Theory and Implications," *Journal of Marketing Research* (31:1), pp 1-14.

Sanders, C. B., and Cuneo, C. J. 2010. "Social Reliability in Qualitative Team Research," *Sociology* (44:2), pp 325-343.

Schuster, D. 2012. "Inter-Coder Agreement for Nominal scales: A Model-Based Approach," *arXiv preprint arXiv:1208.1087*.

Scott, W. A. 1955. "Reliability of Content Analysis: The Case of Nominal Scale Coding," *Public Opinion Quarterly* (19:3), pp 321-325.

Siegel, S., and Castellan, N. J. 1988. "The Case of k Related Samples," *Nonparametric Statistics for Behavioral Sciences. New York: McGraw-Hill*, pp 170-4.

Stevens, M. R., Lyles, W., and Berke, P. R. 2014. "Measuring and Reporting Intercoder Reliability in Plan Quality Evaluation Research," *Journal of Planning Education and Research* (34:1), pp 77-93.

Strauss, A., and Corbin, J. 1990. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Newbury Park, CA: Sage Publications.

Venkatesh, V., Brown, S. A., and Bala, H. 2013. "Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems," *MIS Quarterly* (*37*:1), pp 21-54.

Zhao, X., Liu, J. S., and Deng, K. 2013. "Assumptions Behind Intercoder Reliability Indices," *Annals of the International Communication Association* (36:1), pp 419-480.