Intercoder reliability for qualitative research

You win some, but do you lose some as well?

TRAIL Research School, October 2012

Authors

Niek Mouter, MSc and Diana Vonk Noordegraaf, MSc

Faculty of Technology, Policy and Management, Department of Transport and Logistics, Delft University of Technology, The Netherlands

Contents

Abstract

This paper discusses challenges in testing the reliability of qualitative research. It was found that, especially in the field of transport, it is uncommon to explicitly discuss the reliability of the analysis of interviews and literature reviews. The foundation of these analyses is the coding of interview transcriptions or the reviewed papers using content analysis. The reliability of coding can be assessed through an intercoder reliability check. This paper discusses how this assessment can be carried out. Furthermore it lists lessons learned and gives some practical recommendation based on the hands on experiences of the two authors. This paper could be relevant for all researchers that aim to conduct interviews or to write a literature review and wish to explicitly assess the reliability of their data analysis in order to enhance the research quality of their work.

Keywords

Intercoder reliability, content analysis, qualitative research, transport policy, research quality

1. Introduction

There are several ways to measure scientific quality. Simple measures for scientific quality, such as author quality measured by the Hirsch index, are frequently used in funding, appointment and promotion decisions (Lehmann et al., 2006). However, Lehmann et al. (2006) indicate that such quality assessments based on substantially fewer papers than 50 should be treated with caution. Hence, in these cases other measures for proving research quality are needed.

Qualitative research such as case study research suffers from the common but oversimplified and misleading image that it ‘cannot provide reliable information about a broader class’ (Flyvbjerg 2011, p. 301). This low regard for case study research is at least in part caused by doubt about the reliability of case study findings which are often mistakenly considered less rigorous (Flyvbjerg 2011). Despite these perceptions, there are many measures to ensure and enhance the reliability of case study research. Gibbert et al. (2008) present a framework for methodological rigor of case studies. They list measures with regard to internal validity, construct validity, external validity and reliability. Examples include using a research framework explicitly derived from literature, review of transcripts by peers, explanation of the data analysis and using a case study protocol (see for more examples Yin, (2009)).

This paper focusses on testing the reliability of qualitative research, in particular the analysis of literature reviews and interviews. The reliability of the content analysis, through the coding of the reviewed papers or interview transcriptions, can be tested trough an intercoder reliability check. This paper discusses how to perform this check. Furthermore it lists lessons learned and practical recommendations from the hands on experiences of the two authors. This paper could be particularly relevant for researchers in the field of transport policy that aim to conduct interviews or to write a literature review and wish to explicitly assess the reliability of their data analysis in order to enhance the research quality of their work.

2. Intercoder reliability check

Content analysis has been defined as a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding and categorizing (Weber, 1990). According to Riffe et al. (2005) reliable measurement in content analysis – and in any other research method – is crucial. Without reliable measures any analysis using these measures becomes meaningless. The key words with respect to reliability are transparency and replication (Gibbert et al. 2008). According to Krippendorff (2004) analysts can check the reliability by duplicating their research efforts under various conditions and check the similarities and differences in readings, interpretations, responses to, or uses of given texts or data. For example, by using several researchers with diverse personalities, by working in differing environments, or by relying on different but functionally equal measuring devices. Reliability is indicated by substantial agreement of results among these duplications.

In order to assess the reliability of the coding at least two different researchers must code the same body of content. This intercoder reliability test consists of several main steps:

1) Determine the scope of the intercoder reliability check

Riffe et al. (2005) state that achieving reliability in content analysis begins with defining the categories and subcategories that are most relevant to the study goals and thus must be checked. Often there are time constraints that require a selection of the coding and categorizing process that can be checked by the second coder. This selection can be based on the study goals.

2) Draft the protocol

Definitions and rules that operationalize and demarcate categories and subcategories must be specified in a coding and categorizing protocol. This protocol makes it possible for other researchers to interpret the results and replicate the study. Riffe et al. (2005) state that before carrying out the intercoder reliability test it is essential that the coders are properly trained in using the coding and categorizing protocol in order to be familiar with the definitions of the protocol.

3) Determine the sample that is tested

It is common to limit the intercoder reliability test to a sample of the body of content. Testing all content is impractical and will after a certain point not offer much added value. It depends on the characteristics of the data how large the sample should be. According to, amongst others, Lombard et al., (2004) around 10% of the total content should be sufficient. It depends on the study goals whereas a random or stratified sample is suitable.

4) a) Execute the test, b) select the reliability coefficient and c) calculate the coefficient

The intercoder reliability check consists of coding and comparing the findings of the coders. Reliability coefficients can be used to assess how much the data deviates from perfect reliability. In the literature there is no consensus on a single ‘best’ coefficient to test the intercoder reliability (Lombard et al., 2002). Examples of measurement coefficients that are used in practice are percent agreement, Holsti’s method, Cohen’s kappa, Scott’s pi and Krippendorff’s alpha. The pros and cons of these five coefficients are for instance discussed in Krippendorff et al., (2004), Riffe et al., (2005) and Lombard et al., (2004).

5) Assess the results and draw conclusions

The assessment of results from the intercoder reliability test consists of determining whether the score test is above or below accepted reliability standards for the selected coefficient. However, consensus in literature regarding reliability standards is lacking (Lombard et al., 2002). Neuendorf (2002, p. 145) reviews ‘rules of thumb’ set out by several methodologists and concludes that ‘coefficients of .90 or greater would be acceptable to all, .80 or greater would be acceptable in most situations and below that, there exists great disagreement.’ For instance Riffe et al., (2005) state that a coefficient of .667 would be appropriate for research that is breaking new ground with concepts that are rich in analytical value.

3. Intercoder reliability testing in practice

Before the experiences of the authors with intercoder reliability testing are discussed, first a brief overview is given on some literature searches. Table 1 gives an overview of the search strings focussing on the title, abstract or key words and subsequent hits in the Scopus database for peer-reviewed literature.

Table 1: Overview of search strings and hits in Scopus

Search string	Hits (#)
intercoder reliability	77
intercoder reliability And transport	0
content analysis AND transport	11,358
literature review AND transport	484
Interviews AND transport	1992
content analysis AND reliability AND transport	79
literature review AND reliability AND transport	48
interviews AND reliability AND transport	45

As can be seen in table 1 intercoder reliability only has 77 hits, with more than a quarter being published after 2000. Adding transport to the search string reduced the number of hits to zero. Content analysis, literature review and interviews resulted in subsequently 11.358, 4840 and 1992 hits. When continuing the search in the field of transport with content analysis, literature review and interviews combined with reliability, 79, 48 and 45 hits were found. This is a strong indication that it is not common to perform an intercoder reliability test in the field of transport. In order to reflect on whether this is unjustly or not, the remainder of this section describes how the intercoder reliability for two cases was tested by the authors. The first case concerns analysing interviews about the Dutch cost benefit analysis (CBA) practice and the second case concerns analysing papers for identifying factors that have affected the real-world implementation of road pricing. As at the time of writing the intercoder reliability check of the second case is still on-going, therefore the first case is discussed in more detail. The first case focusses on the Dutch CBA practice. In total 86 key participants in the Netherlands were interviewed via semi-structured in-depth face-to-face interviews. In the interviews, each taking up to 1 to 1,5 hours each, the respondents were asked, amongst others, to mention the five most important substantive problems, the most important advantage and the most important disadvantage they perceive with the appraisal of spatial-infrastructure projects using CBA. For each interview a transcript was made and the content was coded by the first coder. The second case focusses on analysing over 150 papers that discuss the real-world implementations of road pricing. The content analysis will be used in a literature review paper on the success and failure factors for road pricing.

The intercoder reliability of the content analysis of the interviews was tested by an independent coder. In both cases the second coder was familiar in the field of transport policy but did not have case specific knowledge nor was involved in the research. In both cases the level of knowledge of the second coder proved adequate and in both cases the scope of the intercoder reliability check followed the research objectives. It turned out to be too time consuming to test the complete coding process executed by the first coders. The two cases differed in their sampling strategy. In the first case a random sample was drawn from all interviews. In the second case a complete random sample did not suit with the research objectives. In the second case a stratified sample was taken. In addition to checking the identification of success and failure factors in all papers, using a stratified sample also allows for checking whether coding of multiple papers about a specific road pricing scheme (e.g. London congestion charging) would lead to a consistent set of most important factors for that specific case.

To make the coding process transparent and replicable, for both cases a 'coding and categorizing protocol' was developed in which the rules for coding were described. It was found that this protocol was too detailed in the first case and too general in the second case. In the second case the protocol was extended after the discussion regarding the coding of the first paper. The independent coders coded a sample of the interview transcripts and the papers in order to get acquainted with the rules of the 'coding and categorizing protocol'. In total 10 interviews and 10 papers were coded.

After coding the interviews and papers, the coding of the coders were compared and discussed. This step receives relatively few attention in the literature but turned out to be crucial in the final assessment of the reliability of the content analysis. To illustrate the last steps; 4a) carry out the test, 4b) select the reliability coefficient, 4c) calculate the coefficient, 5) assess the results and draw conclusions, only the findings from the first case are included.

To assess the intercoder reliability of coding substantive problems in the interviews the Holsti's coefficient (Holsti, 1969) was selected for two reasons. First, it was not necessary to use a sophisticated coefficient like Krippendorff's Alpha – which is not easy to replicate – because the chance that two coders coded the same quote of a respondent by chance as a substantive problem is considered negligible. Second, Holsti's coefficient gives a more comprehensive insight in the intercoder reliability than percent agreement. The intercoder reliability of the clustering of coded substantive problems in 9 problem clusters was assessed using Krippendorff's Alpha for two reasons. First, this agreement measure considers the possibility that coders cluster a substantive problem into the same problem cluster by chance. Second, Krippendorff's Alpha adjusts – in contrast to, for instance, Scott's Pi – for small sample bias (Krippendorff, 2004; Riffe et al., 2005). The results of the intercoder reliability test regarding these two steps of content analysis are reported in figure 1.

Figure 1 shows that the first coder and the independent coder agreed 47 times that a respondent's quote must be qualified as a substantive problem and 12 times they disagreed. This results in a Holsti's coefficient of 0.886. Moreover, two important causes for disagreement were reported: '*For four of the twelve times that the coders disagreed, one of the two coders coded a quote of a respondent as a substantive problem, whilst the other coder coded the quote as a disadvantage. Five times the first coder coded a quote of a respondent as substantive problem 7 (presentation), whereas the independent coder did not code the quote*'. Moreover, figure 1 shows that Krippendorff's Alpha is .919 for the agreement of the two coders concerning which problem cluster a substantive problem should be grouped under.

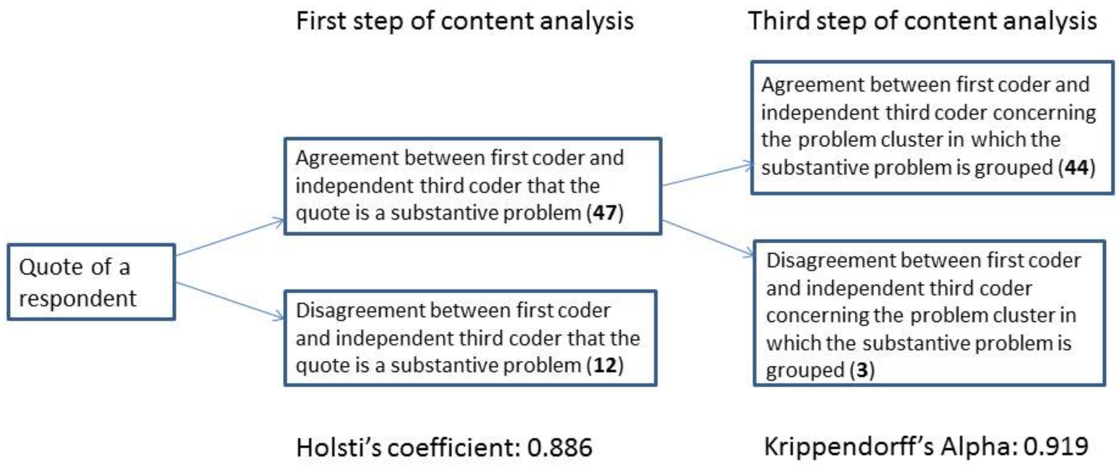
First step of content analysis
Third step of content analysis
Quote of a respondent
Agreement between first coder and independent third coder that the quote is a substantive problem (47)
Disagreement between first coder and independent third coder that the quote is a substantive problem (12)
Agreement between first coder and independent third coder concerning the problem cluster in which the substantive problem is grouped (44)
Disagreement between first coder and independent third coder concerning the problem cluster in which the substantive problem is grouped (3)
Holsti's coefficient: 0.886
Krippendorff's Alpha: 0.919

Figure 1: Results of the intercoder reliability test

Subsequent to reporting the reliability coefficients the paper reported the extent to which it can be determined that the coefficients of .886 (Holsti's coefficient) and .919 (Krippendorff's Alpha) should lead to the conclusion that the content analysis in this study is reliable. Lombard et al., (2002) state that there is no established standard for determining an acceptable level of reliability. Neuendorf (2002, p.145) reviews 'rules of thumb' set out by several methodologists and concludes that 'coefficients of .90 or greater would be acceptable to all, .80 or greater would be acceptable in most situations and below that, there exists great disagreement.' As a result we conclude that the results of the content analysis of substantive problems are highly reliable and not the result of a purely subjective process. Moreover, we reported that there are reliability problems with the number of times problem cluster number 7 is coded by the first coder and as a consequence the frequencies of substantive problems that are grouped into problem cluster 7 should be interpreted with caution.

4. Lessons learned and practical recommendations

The lessons learned and practical recommendations described in this section are based on the experiences of the two authors and limited to two intercoder reliability checks. Although this overview will by no means be exhaustive, it could be helpful to other researchers who also intent to perform an intercoder reliability check on their analysis of their conducted interviews or papers included in a literature review.

4.1 Lessons learned

The most important lesson learned from assessing the scientific quality of the research through an intercoder reliability check is that it makes clear which (preliminary) conclusions can be drawn from the data and, more important, which conclusions can no longer be supported. For instance, in the first case study on CBA, the researcher thought that it was possible to rank the CBA-advantages and CBA-disadvantages. However, after completion of the intercoder reliability check the results proved unreliable and it was concluded that ranking was not valid. Assessing the intercoder reliability check led to drawing more modest results. In fact it even led to including a note to interpret the frequencies of substantive problems of one of the problem categories with caution. Hence, an intercoder reliability check can prevent researchers from “over icing the pudding”. Especially when a researcher has high expectations regarding the research results, it can initially feel as a loss when he finds that the results do not support (preliminary) conclusions. However, we feel it is much more valuable to draw reliable and valid but perhaps more cautious and nuanced conclusions than to put blinders on and simply ignore the limitations in your research.

The second lesson learned is that the discussions on the differences in coding were very fruitful for discovering the causes of the differences. For example it was found that differences in coding were caused by ambiguousness in the text that had to be coded, in the level of detail that each coder interpreted the text (e.g. include or exclude sub factors), by factors for which two categories apply or simply by errors caused by insufficient accuracy in coding. These differences in coding are often related with (unclear) research choices. Hence, once it is clear what causes the differences in the coding outcomes, it becomes much easier to identify and make the limitations of the research explicit.

A third and related lesson learned is that the same discussions on the differences in coding, and in addition to these discussions, the training of the second coder in using the coding and categorizing protocol, were very useful for making methodological choices explicit. The first coder constructed the ‘coding and categorizing protocol’ and therefore the rules of this protocol are straightforward for him. However, the independent coder sometimes does not understand the rules and as a result the first coder is forced to make the rules more explicit. This training process helps to reflect on and defend the methodological choices made in the ‘coding and categorizing protocol’. It can therefore contribute to enhancing the quality of the protocol.

A fourth lesson learned is that it is difficult to obtain the appropriate level of detail of the coding and categorizing protocol. If the coding and categorizing protocol is too detailed it becomes complex and the second coder will not be able to take all decision rules into account, leading to coding errors. For instance, the independent coder did not consider the rule that quotes of respondents concerning the use, the interpretation and the extent to which people understand the CBA are considered as disadvantages during the coding of the last five interviews, which resulted in very low intercoder reliability scores on this aspect. On the other hand, if the coding and categorizing protocol is too generic the second coder has insufficient guidelines to hold on to, which could also result in errors. Hence, the level of detail affects the intercoder reliability scores. This finding is confirmed in the literature (Riffe et al., 2005).

A fifth lesson learned is that performing an intercoder reliability check is very time consuming. It requires a large investment from the independent coder. However, also the first coder has to invest a substantial amount of time for preparing the coding and categorizing protocol, the comparisons of the coding and to jointly discuss the findings.

A sixth lesson learned is that it is easier to carry out a reliable content analysis when the interviews or literature studied discuss exactly the same topic as the topic of your content analysis. More specifically, in the second case the aim was to identify success and failure factors for the implementation of road pricing with content analysis of the literature. In this case it is more likely that analysis of literature that focusses on the implementation of road pricing will be easier to identify and lead to more reliable results than analyzing literature that focusses on discussing the effects of a road pricing scheme or give a general case description.

The last lesson learned concerns a positive external effect of performing an intercoder reliability check. After the check the independent coder has attained in-depth insight in the research of the first coder, making future scientific discussions on this research easier.

4.2 Practical recommendations

A first recommendation is, if possible, to reduce the ambiguousness included in the body of content that needs to be coded. In the first case study it became very clear that the content of the interviews where the interviewer had asked clarifying questions was much easier to code than the content of the interviews where the interviewer did not do this. Hence it is recommended to ask clarifying questions in the interview so that the answers of a respondent can only be interpreted in one way.

A second recommendation for first coders is to keep track of the parts of the body of content that was difficult to code. If the first coder already considers the part of content difficult to analyse, this will surely be confirmed by the second coder. As the first coder gets more experienced with the coding, there will be differences in the coding of the first and later parts of the coding. In the second case, which included over 150 papers, it became very evident that it is impossible to remember which papers were most difficult to code. Therefore, it is recommended that the first coder keeps track of how difficult it was to code each part of the content or which specific aspects caused doubt. Than after finalizing the coding it is recommended to recode the most difficult parts for the content twice and to check the specific parts that created doubt.

A third recommendation, following the third lesson learned, is to do a preliminary ‘feedback’ intercoder reliability check in an early stage of the content analysis – with for instance 3 interviews or papers – in order to discuss and improve the ‘coding and categorizing protocol’. It was found that the discussion with another researcher on the choices made in this protocol can lead to new insights and modification of the protocol. Furthermore, making a coding protocol it is important to tailor this to the needs of the second coder. It is recommended to avoid jargon and to define and explain each concept used in the research (e.g. the categories). To prevent that the first coder has to recode all the content and having to modify the coding protocol, it is recommended to have the discussion on the protocol in an early stage.

A fourth recommendation is to pay attention to the mutual expectations of the first and independent coder on how the coding needs to be executed and how much effort is required. The speed of coding and, related, the level of detail and accurateness of the coding can different tremendously. It is important to discuss these aspects before stating the intercoder reliability check.

Last, it is recommended to make process arrangements between the first and independent coder regarding:

- credits of the work

- the level of influence the independent coder has in modifying the conclusions drawn from the intercoder reliability check
- how to resolve conflicts
- how much time each researcher spends on the check

With regard to the invested time, in the most ideal case the researcher can be each other’s independent coders, as was the case with the two authors of this paper.

5. Synthesis

As we found that in the field of transport it seems very uncommon to check the reliability of the coding of interviews or the review of literature it is safe to say that any attempt to do so is a quick win. In both cases discussed in this paper the authors won in terms of enhanced research quality due to testing the intercoder reliability. The intercoder reliability check made clear to which extent results were reliable. This insight contributed to:

- making methodological choices explicit, and hence it contributed to being able to explain and defend these methodological choices in a more clear and structured manner;
- identifying and explicitly listing the limitations of the research;
- drawing valid conclusions.

We are confident that this has substantially improved the scientific quality of our work. A positive side effect is that the independent coders attained an in-depth insight in the research of the first coder, making future scientific discussions on the research easier.

Question is, did we lose some as a result of an intercoder reliability test? Obviously both coders lost time when performing the test. Especially the time required by the first coder to prepare the coding for the second coder and to jointly discuss the findings should not be under estimated. Furthermore it can initially feel as a loss when the outcomes of the test cause that some (preliminary) conclusions can no longer be supported. However, in our view the loss of aspired results should be celebrated as a scientific victory.

Overall, it is concluded that even in case of a quite extensive intercoder reliability check, the gains outweighed the losses. Therefore performing an intercoder reliability check is highly recommended to all researchers that apply content analysis to interviews or papers included in a literature review.

References

Flyvbjerg, B. (2011). Case Study. The Sage Handbook of Qualitative Research N. K. Denzin and Y. S. Lincoln. Thousands Oaks, CA, Sage: 301-316.

Gibbert, M., W. Ruigrok, et al. (2008). "What passes as a rigorous case study?" Strategic Management Journal 29(13): 1465-1474.

Holsti, O.R., 1969. Content Analysis for the Social Sciences and Humanities. Reading, MA: Addison-Wesley.

Krippendorff, K., 2004. Content analysis: an introduction to its methodology. Sage Publications Ltd. London.

Lehmann, S., Jackson, A. D. and Lautrup, B. E. (2006) 'Measures for measures', Nature, Vol. 444, no. 7122, pp. 1003-1004.

Lombard, M., Snyder-Duch, J., Bracken, C.C., 2004. Practical Resources for Assessing and Reporting Intercoder Reliability in Content Analysis Research. Retrieved April 2008, 2004.

Neuendorf, K.A., 2002. The content analysis guidebook. Thousand Oaks, CA: Sage.

Riffe, D., Lacy, S., Fico, F.G., 2005. Analyzing media messages: Using quantitative content analysis in research. Mahwah, NJ: Lawrence Erlbaum Associates

Weber, R. P. (1990). Basic Content Analysis, 2nd ed. Newbury Park, CA.

Yin, R. K. (2009). Case study research: Design and methods, Sage publications, INC.