Increasing the visibility of coding decisions in team-based qualitative research in nursing

Carl Thompson*, Dorothy McCaughan, Nicky Cullum, Trevor A Sheldon, Pauline Raynor

Room 156, Department of Health Sciences, Alcuin Teaching Building, University of York, York, YO10 5DD, UK

Received 16 December 2002; received in revised form 20 March 2003; accepted 24 March 2003

Abstract


Design: The article presents an illustrative description of the application of the qualitative coding procedure and associated multi-rater Kappa measurement at four time points in 9 months amongst a five person health services research team.

Main outcome measures: The multi-rater Kappa statistic. This is a measure of the extent to which observers achieve possible agreement beyond any agreement expected to occur by chance alone.

Results: Closeness to primary qualitative research data, working relationships over time, and focused research team discussion can all lead to greater agreement and convergence at the level of descriptive coding. The method of measuring agreement between groups of coders was easily applied and appeared a feasible option for similar research projects wishing to demonstrate transparency in their coding procedures.

Conclusion: Measuring agreement beyond chance by the multi-rater Kappa statistic has some utility for research teams whose qualitative coding tasks are primarily descriptive. The method offers a standard and transparent approach for demonstrating agreement between coders and should be a feature of qualitative research reporting where appropriate.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: Multi-rater Kappa; Inter-rater reliability; Qualitative descriptive coding; Collaborative research; Nursing decisions

1. Introduction

Nursing is no stranger to qualitative research and has long been aware of the need to unpack the ‘black box’ of experimental and other forms of evaluation. The profession is equally aware of the importance of the kinds of research questions to which qualitative research is best suited and the need to develop hypotheses for quantitative ‘testing’ which are firmly rooted in the context of real life meaning and settings. In fact, in nursing qualitative research designs dominate (Cullum, 1997). However, alongside the growth in popularity of qualitative research designs, concerns have been raised over the rigour of qualitative research and its susceptibility to bias (Mays and Pope, 1995; Seale, 1999).

Whilst the qualitative research community is often divided in the criteria that should be applied when quality assuring their research products, it is generally agreed that some kind of transparency regarding the analytic process is desirable (Murphy et al., 1998).

Many nursing research studies are collaborative, involving multiple data collectors and team-based approaches to analysis. Such team-based approaches may help to reduce bias associated with individual
interpretation alone but the issue of coding (or assigning meaning to qualitative data) remains problematic. Specifically, the possibility of poor consistency between coders is introduced. Few studies provide transparent evidence of consistency, and most fail to go beyond the (oft seen) generic statement, ‘disagreements in coding were discussed and consensus reached’. When researching the critical appraisal of qualitative research Greenhalgh and Taylor (1997) found no examples of demonstrable inter-observer reliability in any of the qualitative research papers they examined. This article outlines one approach used by a team of researchers to increase the rigour and transparency associated with the assignment of meaning to data in a mixed method research study funded by the UK Medical Research Council.

2. Coding: the common thread in qualitative research designs

Though qualitative research designs differ (for example, ethnographic fieldwork, grounded theory studies and phenomenological enquiry), they all need to interpret the raw data during analysis and convert these into theoretical categories. This process of coding can range from minimally interpretative descriptive coding to interpretative pattern or thematic coding (Miles and Huberman, 1994).

This paper deals with the descriptive level of coding. At this level of coding it is reasonable to assert that agreement between researchers should be a key characteristic of the research. Moreover, given the very descriptive nature of the research task it is both desirable to avoid ‘subjective bias’ and to use or develop coding frames that everyone in the team can utilise.

3. Kappa's contribution to reliability in qualitative research

The idea of Kappa is a simple one: it offers a measure of agreement beyond that which could be expected by chance alone. Such a measure has obvious advantages for those researchers wishing to explore the reliability of the measures employed in the course of a quantitative project. Specifically, it is readily and easily computed, relatively transparent and easily documented, and as long as the measures and processes used do not lead to extreme inter-rater distributions, Kappa offers a reasonably valid measure of reliability. It would be wrong to assume that such advantages are transferred automatically to qualitative research. Whilst there is nothing in the process of calculating Kappa per se that prevents its use in qualitative research, several epistemological and ontological differences need addressing if it is to be used sensitively and with respect for qualitative methods’ theoretical roots. These differences are reflected most vociferously in the debate surrounding the relative merits and demerits surrounding the nature of reliability in qualitative research. Inter-rater reliability is a feature of some qualitative research (Armstrong, 1997; Carey et al., 2000; Roebuck et al., 2001). However, few papers make use of formal Kappa. Where they do, discussion of the place of quantitative measurement of agreement in qualitative studies is rare and cursory (Carey et al., 2000). We have found no other examples of the use of the multi-rater version of Kappa.

Reliability as a mark of quality in qualitative research was largely superseded by the idea of trustworthiness in the 1980s (Morse et al., 2002). Trustworthiness and the closely related idea of rigour has several dimensions: credibility, fit, the presence of an audit trail and confirmability (Guba and Lincoln, 1981). The problem with the idea of trustworthiness is that demonstrating it largely relies on post hoc evaluative procedures on the part of the reader. In other words, the onus is on the reader to apply a series of implicit (and in some case explicit) standards to the paper in order to ascertain the degree of trust they have in its results and conclusions. This post hoc evaluation does not however guarantee the quality of the research, and as Morse and colleagues point out, has little to do with the actual attainment of reliability — and for that matter validity (Morse et al., 2002). Two of the most common cues intended to steer the reader to a diagnosis of trustworthy qualitative research are the presence of an audit trail and member checking. Morse and colleagues (2002) highlight how simply being able to demonstrate decision points in a project (via an audit trail) provide the reader with few clues as to the quality of those decisions. Similarly, member checking applied blindly is a similarly flawed measure of quality. Qualitative research (with the possible exception of case study) often relies on synthesis of accounts, a degree of necessary decontextualisation, and abstraction from and across individuals. Given these characteristics why do we expect participants to be able to recognise themselves or their experiences? Based on these examples Morse et al. (2002) argue that sections of the qualitative research community have failed to make the distinction between ‘procedures that determine validity during the course of inquiry, and those that provide the research with credentials on completion of the project… the terms validity and reliability remain pertinent in qualitative inquiry and should be maintained’ (Morse et al., 2002). Clearly then, verification strategies need to be built into the design and process of the study.

It is with this in mind that we undertook to use such verification strategies to shape our enquiry. Whilst mindful of the need to be sensitive to the paradigm in which the study was located it was felt that the added value of such an approach to the process of analysis was
worthwhile. Alternatives to using Kappa to examine inter-rater reliability, such as the qualitative technique of having a third party examine the coding choices made by team members as used by Roebuck et al. (2001) were another option. However, as we had five members of the team involved in the coding process already we were unsure how such strategies would add any extra value to the analysis or verification of choices.

4. The study context

The study to which this method was applied was a mixed method, multi site, descriptive–analytical case study examining the use of research based information by primary care nurses. This built on an earlier study examining research information use by acute care nurses (Thompson et al., 2001). A key research question associated with both studies was, “what kinds of decisions do nurses make?” We felt that if we knew the answer to this question then the research community would be better placed to match nurses to the kinds of research-based knowledge that would best address the uncertainties associated with their clinical decisions.

5. Methods

Early in the project the research team agreed the conceptual categories associated with a typology of kinds of decisions. This agreement was based on the scrutiny of early pilot and case site interview transcripts and field notes (see Table 1). This was done by team-based discussion and revision. As the project developed the principal investigator for the project (CT) scrutinised the interview transcripts and observational field notes and descriptively coded (as clinical decisions) those sections of the text in which nurses were faced with choosing one or more discrete options. Following the construction of the typology we were keen to ascertain (a) how well the typology ‘fitted’ the real world data still to be collected — i.e. its construct validity, (b) whether other researchers, less close to the raw data, would also code the data in similar ways, i.e. was the typology a function of closeness to data collection (c) could we make the variability in coding between researchers the basis for refinement and adjustment of the developed decision typology and if so how?

Answering these questions required the descriptive coding of interview transcripts (where people discussed decisions made) and observational field notes (where clinical choices were observed) (Thompson et al., 2001).

CT presented a selection of 50 decisions to the immediate research team (DM and PR) on four occasions (a different set of 50 decisions at time 1, the beginning of main data collection; time 2, mid point; time 3, end of main data collection and time 4, after discussion and refinement of typology with rest of the team). TS and NC were presented with the decisions at times 3 and 4. Team members were asked to classify the decisions into one of the 13 categories in the typology. The working hypothesis behind presenting the decisions to TS and NC at time period 3 would be that there would be less agreement between them and the rest of the team as they were ‘further’ from the data and the analysis than the other three researchers (CT, DM and PR). The working hypothesis behind presenting the data at time 4 would be that the levels of agreement between all reviewers would improve following discussion, clarification and refinement of the coding frame (decision typology). After time 3, discussion took place and consensus was reached regarding the most contested or troublesome elements of the typology. CT, PR, DM and NC were present at this discussion, TS was not. Our hypothesis being that whilst the gap between NC and the 3 main researchers on the project would narrow at time 4 the gap between TS and CT, DM and PR would not.

The resulting decision codes x raters data matrix was analysed using SPSS (version 10). We used the multirater Kappa measure outlined by Siegel and Castellan (1988) and the SPSS ‘include’ feature. Both are available for free download from http://www.spss.com/tech/stat/macos/MKappasc.html.

Kappa measures the extent to which observers achieve the possible agreement beyond any agreement expected to occur by chance alone. Kappa values can be interpreted thus: 0 = poor agreement; 0–0.2 = slight agreement; 0.2–0.4 = fair agreement; 0.4–0.6 = moderate agreement; 0.6–0.8 = substantial agreement and values of 0.8–1.0-almost perfect agreement (JAMA, 2002).

6. Results

The Kappa scores and comparisons for time periods 1–4 are presented in Table 2. Table 2 shows that the
Kappa scores at the beginning of the project for groups of raters (grouped by working relationship) were noticeably lower than at any of the other time points. The overall trends however appear to show improvement in the amounts of agreement reached and a degree of convergence between coders; although in some case (DM and PR) there is broad fluctuation (0.5, 0.89, 0.65, 0.83). As DM and PR were the primary data collectors we would have expected them to have broader agreement at time 1. However, PR was new to the team and the project’s analytical framework, whereas the other team members had all worked together on the previous (and similar) NHS R&D project.

7. Discussion

The results show that despite agreeing the broad conceptual categories in the coding frame at an early stage in the project the coding process was still a largely subjective affair with varying amounts of convergence and divergence within groups of coders and as a team. As codes were discussed, and disagreement resolved, agreement between raters improved. It is worth noting that that the agreement between TS and the 3 ‘core’ researchers (CT, PR and DM) whilst still adequate at time 4 was lower than the agreement between NC and the same core researchers. Only NC was present at the revision of the codes at time 3, whilst TS was simply aware that the codes had been discussed and disagreements resolved. As expected, people who had worked together longest (CT & DM) had closer agreement than individuals who were new to the team (CT & PR for example)—this approach to measuring agreement simply exposes that relationship.

8. Limitations of the approach

Despite proving useful, the procedure was not without its limitations. The technique is only suitable for data at the most descriptive level of coding. For some studies this would be a major drawback. An example is detailed ethnography, which relies heavily on interpretative coding and ongoing and the reflexive data analysis of a single field researcher. Even where such studies have a team component to validation of findings or in reviewing results, the team members’ distance from the data, and the heavily interpretative nature of the assignment of meaning to data, probably precludes the use of multiple coding strategies. Of course, the converse of this is that for studies based around content analytic designs the procedure could be eminently suitable.

Moreover, we would have liked to repeat the exercise over enough time periods to enable the statistical testing of trends to ensure that the apparent convergence of coding patterns and raised levels of agreement were not just due to chance. A point other research teams may wish to consider.

Kappa has several undesirable statistical characteristics (Cicchetti and Feinstein, 1990; Feinstein and Cicchetti, 1990). Specifically, when the distribution of decisions being measured becomes extreme then possible agreement beyond chance (which is what Kappa is measuring) becomes hard to achieve. Consequently, even though agreement may be high the Kappa values may be low. Interpretation is therefore complicated by the nature of that which judges are being asked to agree on. There are a number of examples of this phenomenon in healthcare (Cook, 1990). There are statistical techniques that consider the dependence between raters. These

Table 2
Multi-rater Kappa scores (and standard errors) for team over four time periods

<table>
<thead>
<tr>
<th>Rater comparison</th>
<th>Time 1 start of data collection (SE)</th>
<th>Time 2 mid point (SE)</th>
<th>Time 3 end of data collection (SE)</th>
<th>Time 4 post data collection (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT, DM, PR, TS, NC (1a&amp;b, 2)</td>
<td>—</td>
<td>—</td>
<td>0.62 (0.03)</td>
<td>0.80 (0.02)</td>
</tr>
<tr>
<td>CT, DM, PR (1a,1a,1b)</td>
<td>0.63 (0.05)</td>
<td>0.81 (0.04)</td>
<td>0.71 (0.03)</td>
<td>0.86 (0.04)</td>
</tr>
<tr>
<td>CT, DM (1a,1a)</td>
<td>0.79 (0.07)</td>
<td>0.81 (0.06)</td>
<td>0.85 (0.06)</td>
<td>0.96 (0.06)</td>
</tr>
<tr>
<td>CT, PR (1a, 1b)</td>
<td>0.59 (0.07)</td>
<td>0.73 (0.07)</td>
<td>0.60 (0.07)</td>
<td>0.79 (0.06)</td>
</tr>
<tr>
<td>DM, PR (1a, 1b)</td>
<td>0.50 (0.07)</td>
<td>0.89 (0.07)</td>
<td>0.65 (0.06)</td>
<td>0.83 (0.06)</td>
</tr>
<tr>
<td>NC, TS (2, 2)</td>
<td>—</td>
<td>—</td>
<td>0.63 (0.05)</td>
<td>0.78 (0.06)</td>
</tr>
<tr>
<td>CT, TS (1a, 2)</td>
<td>—</td>
<td>—</td>
<td>0.59 (0.06)</td>
<td>0.73 (0.06)</td>
</tr>
<tr>
<td>CT, NC (1a, 2)</td>
<td>—</td>
<td>—</td>
<td>0.59 (0.06)</td>
<td>0.84 (0.07)</td>
</tr>
<tr>
<td>DM, TS (1a, 2)</td>
<td>—</td>
<td>—</td>
<td>0.58 (0.08)</td>
<td>0.71 (0.07)</td>
</tr>
<tr>
<td>DM, NC (1a, 2)</td>
<td>—</td>
<td>—</td>
<td>0.61 (0.06)</td>
<td>0.85 (0.07)</td>
</tr>
<tr>
<td>PR, TS (1b, 2)</td>
<td>—</td>
<td>—</td>
<td>0.48 (0.06)</td>
<td>0.75 (0.06)</td>
</tr>
<tr>
<td>PR, NC (1b, 2)</td>
<td>—</td>
<td>—</td>
<td>0.48 (0.06)</td>
<td>0.75 (0.06)</td>
</tr>
</tbody>
</table>

Note: The numbers in parentheses indicate the ‘distance’ from the data of the coders and therefore constitute the working hypotheses. The scale ranges from 1a to 2; with 1a being closest (at the time the research was undertaken) and 2 the furthest. So (1a, 2) constitutes the hypothesis close versus furthest therefore less agreement beyond chance. There is no agreed procedure for generating confidence intervals for the multi-rater version of Kappa, therefore only associated standard errors are presented.
involve modelling of rater and object (in this case the decisions) effects and could be used to overcome the limitations of Kappa. Alternatively, Cicchetti and Feinstein (1990) suggest generating separate indexes of proportionate agreement in the observers’ positive and negative decisions as a way of considering the sensitivity and specificity of rater decisions. However, our need was for a ‘quick and dirty’ tool to simply raise the visibility of the decisions made within the team over time and both of these techniques would require considerable statistical effort when applied to modelling trends between raters. The standard errors presented with the Kappa scores in Table 2 suggest that such approaches might prove useful for future work.

For studies such as ours, with multiple research questions—some of which require only limited interpretation of data, and interpretation likely to be shared by a number of researchers—then the multi rater Kappa procedure appears useful. Specifically, it offered a focus for discussion and scrutiny of differences; for example, the categories of assessment and diagnosis were subject to variability in coding in our coding frame. The exercise forced us to clarify what we meant by assessment in the context of clinical decision making. We characterised it as the choice to assess or not to assess, or the choice of what to include in the assessment or not. Assessment (as a process) we classed as clinical judgement (i.e. the weighing up of information). Diagnosis was the assignment of the patient to a formal diagnostic category such as chest infection.

The other primary benefit of this approach is that it allows team members to play an active role in the analysis of data. This quantitative indicator of agreement with respect to coding between and within rater groups complements the more qualitative impressionistic judgement of ‘trust’ that one places on team members’ roles in analysis. Moreover, the technique appears sensitive to clarification and revision of coding definitions within the team.

Studies such as those conducted by Ely and colleagues (Ely et al., 1999) examining the clinical questions asked by primary care doctors, and utilising a similar level of coding use conventional two-rater Kappa scores. The approach of Siegal and Castellan (1988) has the advantage of being able to calculate Kappa for more than two raters and hence promote exploration of team-agreement.

Other computerised data management packages such as QSR’s Nvivo Merge (http://www.qsr.com.au/products/mfnvivo.html) have the ability to evaluate consistency in coding between multiple team members. However, our approach meant that team members unfamiliar with such software, and the proprietary nature of the software itself, could play a fuller role in analysis than would otherwise have been the case.

9. Conclusion

The multi-rater Kappa procedure outlined in this article can prove useful for team based qualitative nursing research projects in which coding tasks are not dependent on detailed interpretation. The technique quantifies agreement between more than two researchers and can usefully act as the basis for discussion of differences and improvement (convergence) of coding processes. Moreover, multi-rater Kappa can help in the refining and adaptation of coding frames and ultimately in the quality of any final analysis. We feel that the procedure offers a standard approach that other researchers may wish to consider using and reporting when disseminating their qualitative research findings. However, depending on the nature of the agreement distributions and the scale of the research task, teams may need to consider the limitations of the Kappa statistic and the techniques of overcoming these. Despite the technique’s—and our study’s limitations—we believe that the approach has considerable benefits over the traditional method of assuring the reader of the probity of coding decisions taken in team-based qualitative research: ‘trust us, we talked about it and resolved the disagreements’.

Acknowledgements

The UK Medical Research Council funded the study underpinning this article. The earlier study mentioned was funded by the UK NHS Research and Development Programme and led to the current study being conducted and the need to demonstrate reliability in coding procedures.

References


