‘Opening Up Open-ended Survey Data Using Qualitative Software’

Quality & Quantity

October 2013, Volume 47, Issue 6, pp 3261-3276

Jane Fielding, Nigel Fielding and Graham Hughes

Department of Sociology

University of Surrey

Guildford GU2 7XH

Abstract

This article considers the contribution that qualitative software can make to ‘opening up’ Open-Ended Question (‘OEQ’) data from surveys. While integrating OEQ data with the analysis of fixed response items is a challenge, it is also an endeavour for which qualitative software offers considerable support. For survey researchers who wish to derive more analytic value from OEQ data, qualitative software can be a useful resource. We profile the systematic use of qualitative software for such purposes, and the procedures and practical considerations involved. The discussion is illustrated by examples derived from a survey dataset relating to environmental risk in the UK.

Keywords

OPEN ENDED QUESTIONS

SURVEY METHODS

QUALITATIVE SOFTWARE

Integrating Open Ended Question ('OEQ') data with the analysis of closed response items is a longstanding challenge in survey research. This article profiles the contribution that qualitative software can make to 'opening up' open-ended survey data and offers guidance to those interested in maximising the analytic return from such data. We present computational strategies for better integration of data from OEQs with closed response items in surveys and the principal procedures and practical considerations involved.

OEQs may be included in social survey instruments to capture dimensions not represented in the numerical items or anticipated by the survey designer, to enable respondents to provide reasons or background informing the option they have selected from the response set provided in fixed response questions, or even to make respondents feel that the survey instrument has a human and less constraining face. However, when the time comes for analysis, the survey researcher may find that the OEQ data are as much an irritant as a resource. Compared to the closed response questions whose analysis can be automated with a statistics package, the OEQs are out of step. The survey analyst is likely to endorse Miles' (1979) classic characterization of qualitative data as 'an attractive nuisance'.

Yet the custodians of prominent surveys such as the American National Election Survey ('ANES') and the British Crime Survey recognize the need to more fully and systematically exploit OEQs rather than skimming OEQs haphazardly in search of juicy quotes to flesh out a statistical relationship. Declining survey response rates, the increasing cost of fieldwork, and major institutional investment in data archiving on an international scale, all argue for maximizing the return from existing data sets. For the epistemologically minded, interrelating findings from OEQs and fixed responses elicited at the same time in the same instrument and around precisely the same topic may also overcome a principal

objection to methodological combination, that of paradigmatic (in)commensurability (for example, the argument that different methods inescapably measure different things; see N. Fielding 2009).

Stories of studies that end up not analyzing, transcribing, or even accurately recording response to OEQs continue to be encountered. We believe that Computer Assisted Qualitative Data AnalysiS ('CAQDAS') can be a useful resource in such a context. Our discussion is aimed at the experienced survey researcher with an interest in the possible contribution of CAQDAS to the analysis of OEQs but who does not have any great degree of prior acquaintance with qualitative software. There are around 20 CAQDAS packages on the market, with around 5 market leaders. This article does not set out to provide a comparative review, but for such information see the comprehensive resources at the CAQDAS Networking Project and Online QDA web sites (in the list of web sites at the end of this article). In this article we will discuss generic procedures, occasionally noting package-specific features where these are distinctive.

When interrelating closed response and OEQ data using CAQDAS software, the researcher will be able to:

- More easily automate the process of coding textual responses
- Move instantly between analytic expression and underlying data – explore the data in context
- Form collections of response to OEQs by theme or respondent characteristics
- Make use of textual data visualization features akin to those used for statistical visualization

- Employ selective retrieval strategies for systematic analysis and comparative work
- And maintain an analytic 'audit trail'- increase transparency and conduct checks for validity and reliability.

Open-ended survey data as an established and growing issue

The methodological literature suggests that the management and analysis of OEQ data from surveys is a long-established concern (Frisbie and Sudman 1968, Montgomery and Crittenden 1977) and poses some fairly intractable problems (Giorgetti and Sebastiani 2003). Even where survey research is not pressured by a sponsor wanting results quickly, OEQ data call for extra time investment. These concerns are attested by efforts by the custodians of major, long-term surveys to identify ways to extract more value from OEQs without unduly impeding the timely analysis of the survey. For instance, in December 2008 researchers responsible for the ANES convened a meeting of social research methodologists to consider different approaches to fuller exploitation of OEQ data, most of which featured some degree of automation. A dramatic indication of the longstanding nature of the issue was one speaker's reference to the major current threat to ANES OEQ data. This was not a reference to the usual controversies about the status of qualitative data but rather to rodent infestation in a warehouse in which older OEQ data were stored. The survey, which US law requires to be carried out following every general election and which is formally sponsored by the US Congress, began in 1948. Some OEQ materials had already been lost. The same speaker observed that OEQs had been a bone of contention since at least the 1960s. Some directors of the survey (which passes from one group of academics to another after each sweep) had not much valued the OEQ data but latterly it had been recognised as an underutilised resource.

We anticipate that such issues are likely to grow due to the increasing availability of free or low cost online survey tools (e.g., SurveyMonkey, KeySurvey, and Survey Said). The strapline for KeySurvey makes the point clearly: 'Survey software everyone can use'. It is apparent from graduate research methods teaching that students are enthusiastic adopters of such software, and it is also in prominent use in surveys conducted by community organisations and other voluntary sector users. These packages can produce satisfactory quantified results and presentations, but they provide no more support for OEQ analysis than professional statistics applications used by survey researchers. Yet it is likely that their users will generate OEQs, because asking people to respond in their own terms is as appealing now as ever, and may indeed resonate with the open, democratic qualities associated with the Internet. There may thus be an impetus to include OEQs at the design stage without much consideration of how to handle them analytically. The users of online survey tools may find their OEQ data as anomalous and awkward as did previous generations of survey researchers.

That the issue of handling OEQ data remains current is evident in online methodology forums, where discussions include fully automating the coding of OEQs, indicating the troublesome nature of manually coding large volumes of OEQ responses. For this reason we largely focus our discussion of analytic strategies in this article on the semi-automation functions in CAQDAS. Survey researchers' concerns are not confined to the time that coding OEQs takes. Coding is an additional potential source of measurement error. CAQDAS packages offer tools to track measurement error arising from problems with inter- or intra-coder reliability. In our assessment, CAQDAS offers a helpful middle path with partial automation that allows researchers to pay due attention to the detailed terms and context of OEQ responses.

The exemplar survey

Our discussion illustrates the use of qualitative software applied to an exemplar dataset arising from a professionally-administered survey of risk-related attitudes of members of the UK general public who had experienced flood emergencies. The original study was conducted by a large well established survey organization for the UK Environment Agency, and was made available to academic researchers performing follow-up research for that sponsor (see J. Fielding et al 2005).

The 2001 Post Event Survey was carried out for the UK Environment Agency during June and July 2001 following the serious flood events in the UK during the previous autumn. Such Post Event surveys are carried out on an *ad hoc* basis as and when a flood event occurs. The objective of these surveys is to assess the efficiency of the flood warnings service. For the 2001 survey, a total of 1,257 face to face interviews were conducted in twelve areas in the UK in residents' homes using CAPI (computer-assisted personal interviewing). The average response rate in the twelve areas was 72%.

The following table displays the question wording of selected relevant verbatim responses collected during this survey and used in this article. The OEQ questions in Table 1 were asked only of those respondents who had received a prior warning of any kind regardless of whether they were flooded or not. The unprompted OEQ advice question (Q40) was followed by a prompted advice question (Q41). Questions 57 and 58 were only asked of those respondents who had previously indicated in Q41 that they had remembered being given advice to use sandbags (even if they had not acted on that advice).

[Table 1 here]

Table 1 Open-ended questions in the Post Events survey

CAQDAS and Open Ended survey data

A first consideration is how to organize the data to support the intended analysis. Most qualitative analysis starts with examining all of the data collected from one case (for example reading the full transcript of the first interview), whereas initial quantitative analysis tends to focus on the set of responses by all cases to a single question (for example the variable of interest). When working with open ended questions asked in a survey situation the analyst will have to make a decision whether to approach this data from the perspective of the variable (see Table 2 below) or the case (see Table 3 below). Some CAQDAS software forces this decision to be made at the data preparation stage, as the responses have to be grouped in discrete documents for use within the program, while other software permits the data to be displayed in either way, thus allowing the decision to be deferred until the start of analysis. Whenever the decision is made, it is important that the analyst makes a considered choice as to which approach to use on the basis of analytical requirements rather than habit.

[Table 2 and Table 3 about here]

Table 2: Layout of the responses from the perspective of the variable/question

Table 3: Layout of responses from the perspective of the case/respondent

Data preparation issues

A mundane but essential element to include in all data preparation is a unique respondent identifier attached to all responses and case attributes in order to facilitate correct matching both within and outside the CAQDAS program. Consideration should then be given to exactly what data elements will be required during the analysis process. These will usually include socio-demographic variables such as age and gender, but will also include other factors pertinent to the research topic (for instance, in our worked example on experiences of flooding it was useful to include two variables which recorded the extent to which the respondent's home had been flooded and whether they had received a warning before the water arrived). Different CAQDAS packages have different requirements for how these variables are introduced and manipulated so it is worth spending time at the start of the process identifying the important analytic variables to include (see package-specific Working Papers on the CAQDAS website).

Importing the data into CAQDAS software from a statistical survey package is usually straightforward. It is a matter of selecting the appropriate variables, extracting them to an intermediate format such as a textfile or spreadsheet, and then importing that file in the manner required by the CAQDAS program, in most cases in an MS Excel format. However it should be noted that the attribute or variable labels are more useful in CAQDAS than the numerical values used by statistical packages, and scaled or interval-level variables (such as age) may be more easily used in CAQDAS if they are converted to ordinal variables (such as age-ranges by decade). It should be noted in this context that when saving data in MS Excel format from IBM SPSS, one can select that value labels (where they are defined) can be saved to the spreadsheet instead of data values.

Coding

Coding is a fundamental activity in both survey research and qualitative analysis. In most quantitative surveys the coding 'work' is done by the respondent in the sense that a choice of response categories for each question is provided by the questionnaire designer (for example in the form of a showcard) and the respondent is asked to select the one that appears to be closest to their answer. With the analysis of OEQs this work may be seen as shifting to the researcher post-survey.

Higgins et al (1996) took one minute to code each half line of text for qualitative analysis, using the (now obsolete) NUD*IST package. A survey of 500 respondents using four OEQs would be 1000 lines, or 2000 minutes, equating to 39 hours or roughly one working week. In fields like market research that is probably too slow, but for academic or policy research it is more likely to be an acceptable time investment, and coding in CAQDAS packages has in any case moved on considerably since Higgins et al's estimate, with more automation.

The first task for the analyst is the construction of a codeframe, or a set of coding categories, which can be applied to the data. In qualitative work this is often seen as an iterative process, with ideas and themes being drawn from interpretations of some of the data, refined in the context of inspecting further data, and then applied to all of the data. The refinements may take the form of splitting categories into those with finer distinctions, or involve combining categories together as deeper themes become apparent. Here the ability of CAQDAS to instantly re-code data segments selectively or *in toto* is helpful.

Where the data consist of a large number of brief responses, or where there is evidence of mediation (for example because face-to-face or telephone interviewers have

paraphrased or edited the responses while typing them), it may be appropriate to use computing tools to identify the most common words or phrases and apply coding automatically to the responses that include these terms. In particular, in cases where the researcher wishes to extract quantitative data of the form ‘X% of responses mentioned concept Y’, perhaps to get a sense of how opinion is distributed amongst those who chose to provide OEQ responses and thus to compare it to the distribution of opinion in the overall survey sample on the associated closed response question, a degree of automation may be helpful and justified. Here three types of computer tool may prove useful: word frequency calculators, text searching functions, and autocoding facilities. Use of these is not an abdication of the researcher’s responsibilities because deciding which frequently-used words should be grouped together and what code label should be attached to those groups, remains critical. An advantage is that such an approach may be more transparently rooted in the data and less prone to researcher bias than wholly manual coding. Where autocoding by selected keywords is to be used, the word frequency tool is also useful for identifying the range of variations and misspellings actually present in the data, and can thus significantly improve the accuracy of such a procedure.

A suggested method of working would be to run a word frequency analysis on the set of responses to a question and then, using that output in conjunction with text searches or keyword-in-context functions, to identify recurring themes in that data. An iterative approach may be beneficial as potentially useful words are tested through the careful examination of a sample of their appearances in the data. When the main themes and their main indicator words or phrases have been identified a code can be set up for each and applied to the full set

of responses using autocoding routines. This is illustrated in a schematic diagram in Figure 1 below.

[Figure 1 about here]

Figure 1: An Iterative Approach to Thematic Coding

The key step in Figure 1 is the decision, following a text search or keyword-in-context report, as to whether a selected word is useful or not. ‘Useful’ words will occur with quite high frequency (indicating some commonality amongst respondents) but not so frequently that they lack any power to discriminate between respondents. They will have a link of some sort to the topic under investigation and they should not have too many separate distinct meanings (to avoid ambiguity). Similarly, words that prove to be ‘not useful’ are likely to be those that have many possible meanings and uses (in which cases the other words around them that indicate the sense in which they have been used may prove to be more helpful for analysis), or are words that cover matters not relevant to the current analysis.

In Figure 2 an output from a CAQDAS package is shown. Here a Word Frequency query has been run on responses to a question about how people might be warned of an impending flood, and the results have been displayed in the form of a ‘Tag Cloud’. In Figure 2 the size of font used to display a word is proportional to its frequency in the data, so the most frequently occurring words appear the most prominent. From this display alone, without reading any of the data, one may begin to form ideas about themes in the responses. Some care is needed when interpreting a word frequency output as variations in the spelling of a word can split its count between several ‘hits’. In Figure 2 it may be noted how ‘house’,

‘loudspeaker’ and to a lesser extent ‘phone’ are displayed less prominently than they would be if their plural versions were included with the singular. In conventional text analysis one would use stemming or lemmatization functions to simplify the table, but when the output is subsequently to be used in autocoding processes it may be better to identify the actual spelling variations present in the data and so such functions are less relevant here.

[Figure 2 about here]

Figure 2: NVivo Word Frequency Tag Cloud Output

Semantically, the next step is to examine frequently-occurring phrases. In Figure 3 another CAQDAS output demonstrates how a ‘Phrase finder’ function can augment the word frequency facility. Here data from the survey of people involved in flooding incidents have been analysed to identify the most frequently used phrases in their answers to a question about why they did not use local radio broadcasts to obtain information about the flood. This output may give even more insight into the dataset being analysed. Such relatively limited forms of quantification tell researchers something about patterns of response without stripping context away or violating the interpretative precepts and constraints of qualitative methodology, and qualitative software offers features enabling researchers to get quickly back to context when assessing the meaning of the simple content analysis counts.

[Figure 3 about here]

Figure 3: QDA Miner, WordStat module, Phrase finder output

CAQDAS programs generally allow some flexibility over how much text should be included in the coded segment during autocoding procedures. Where the responses are generally short

it can be useful to include the whole paragraph so that the main context is shown in direct output listings for a code.

A valuable feature in CAQDAS packages is the option to store a definition for each code. When several autocoding runs have been performed using different words or variations of spelling, this can be extremely useful as the place for an accurate record of those parameters. When this is done, the analysis is replicable and transparent, something that is not always so when statistical programs or spreadsheets are used for this task.

As with any coding procedure, there is the need to apply rigorous checks to confirm the accuracy of the coding, irrespective of the method by which coding has been done. The analyst needs to be confident that each application of the code is closely equivalent in meaning, and that no other responses have been omitted from that code despite carrying an equivalent meaning. CAQDAS packages readily list all the passages that share a particular code, and this would assist with the first consistency check. Checking for omissions involves excluding all the passages that have been allocated a particular code and then searching the remainder for similar meanings. Accuracy may be further improved when such consistency checks use a different method to that applied in the main coding work. Thus codes that have been applied one instance at a time by a coder exercising judgement may be checked with computer operations, while the results of autocoding procedures may be checked by human interpretation of samples drawn from the coded responses.

In the process of coding it is likely the researcher will find some responses that are particularly vivid and a few may contain unusual or original ideas. These can be difficult to place in a table of results but could be valuable outputs of the study. In our example we were

struck by the following comments in response to a question about better ways to warn people of an imminent flood:

> *When police are used and or army there should be local back up to provide accurate and more local knowledge especially with regard location. My personal experience was being told my family had been evacuated when they had not, and not being allowed access to my own road when my family were trapped inside.[122284]*

This highlights how authorities may incorrectly assume that the only people affected by a flood are those currently within the flooded area, a unique but valuable observation. Further it was an observation that was not apparent from the closed questions in this survey, which did not include a question about the whereabouts of all members of the household during the flooding event.

Support for Analysis

The assistance that CAQDAS programs can provide for the analysis of OEQs goes well beyond autocoding by keywords. An important element in qualitative analysis is that of keeping in close touch with the full response data throughout the process, so that emerging ideas or theories can be tested against the original words in context. This facility can be extremely useful for extracting all of the responses to a particular question made by a subset of respondents who share a specified combination of attributes. For example, it was revealing to compare the responses to a question about methods of sending out warnings of imminent flooding between those who had been seriously flooded and those who had not been flooded at all. Alternatively, it is also possible to display in another part of the CAQDAS program

screen the attribute set for the respondent whose data is being read, so that their responses may be interpreted in the light of any of those attributes.

In Figure 4 a working screen from a CAQDAS package shows how context can be maintained between different kinds of data during analysis. Five interdependent panels are visible. The 'Document system' (top left) holds the respondent identifiers, the 'Text Browser' (top centre) displays all of the responses by one respondent (number 34304 in this example), the Code System (bottom left) shows the hierarchical structure of codes, the 'Retrieved Segments' (bottom centre) displays all of the responses for a selected combination of codes and respondents, and the 'Variables' (right) shows the data for selected variables ordered by respondent. The data displayed in these windows are linked interactively so, by clicking on a response in the Retrieved Segments window, the user can see that response within the context of all of the other answers given by that particular respondent (in the Text Browser), and the attributes of that respondent are highlighted in the window to the right. Clicking on another response in the Retrieved Segments causes analogous data for another respondent to be displayed in each of the other windows.

[Figure 4 about here]

Figure 4: Working Screen Layout in MAXQDA

In the example shown in Figure 4 a filter has been applied so that only the responses of people aged 55 years and over who mentioned evacuation from their homes because of the flooding are shown in the scrollable Retrieved Segments window. This would enable the researcher to compare responses amongst a group whose age may make them especially vulnerable in a flood emergency.

However, as discussed above, time constraints or data volumes may limit the opportunity to keep on returning to the source data for interpretation. Once thematic coding has been completed and checked it will be possible to search for patterns of responses by using the code frequencies instead of the detailed data. Generally CAQDAS packages include sophisticated cross-tabulation functions allowing the analyst to generate tables based on selected combinations of thematic codes and attributes. These can be switched between counts and percentages and between various bases of calculation to assist with pattern searches. Some packages allow colour shading in the style of heatmaps to add visual emphasis to these tables. For those with reservations about these procedures when applied to qualitative data, it should be noted that it is always possible to jump directly to the subset of responses that match a selected cell in the table, keeping the full context of an observed pattern in close reach.

In Figure 5 an example from the flooding survey data shows a summary of the coded words for the advice that respondents remembered receiving before their flooding episode, plotted against the severity of the flooding that they actually experienced. Those who were not actually flooded claim to have been advised about sandbags much more than those whose houses had been flooded. ‘Sandbags’ is the second most frequent code for the not-flooded subgroup ‘Notfl’ but is equal sixth in frequency for the seriously flooded sub-group. Further analysis examining the incidence of the sandbags code across the towns surveyed suggest that this is not a result of different advice being given in different places, as this code was applied to responses from people in ten of the twelve locations. It facilitates analysis of such surprising findings that clicking on any cell in the Crosstabs output will bring up that set of coded passages, each of which can be viewed in the full context of all the responses by that

respondent. Alternatively, by filtering according to the relevant attributes it is a simple matter to output each set of code/attribute combinations and examine these closely to detect differences or patterns in the language used. In addition, this crosstabs table can be displayed using either row or column percentages of response.

Thus while it may seem anomalous that only 4% of flood victims commented that they remembered being advised to use sandbags compared to 13% of non-flooded respondents, there may be an explanation within the other OEQ data and its interplay with the closed response data. Further analysis of another OEQ asking why sandbags were not actually used to block air vents showed that flood victims believed them to be ineffective. Comments such as ‘can't hold back the tide’ [88911] or more pragmatically ‘does not work with my property as it is on a gravel base’ [12260] were offered. This was in contrast to the responses from the non-flooded respondents to the same OEQ who commented that they either didn’t need to use sandbags because they were not in any danger, or were not provided with sandbags (eg. ‘Considered that there was no risk’ [19503] and ‘no sandbags available at this time all taken up’ [34304] and ‘no sandbags available as most went to each end of the village’ [34321]). Of those who suffered the most severe flooding in autumn 2001, 51% had previous flood experience in their current home compared to 2.4% of those who had never been flooded before. Thus it seems that previous flood victims responding to an OEQ asking what advice they remember receiving before the flood, may be responding in light of their *previous* flood experience. This may explain why the advice to use sandbags (which their experience showed not to be effective) was not the advice they thought of when asked to respond to the 2001 post experience survey.

[Figure 5 about here]

Figure 5: MAXQDA Crosstabs output for remembered advice and extent of flooding

Whichever of these techniques or functions is used, the essential benefit of using a CAQDAS program to carry out the analysis is that the original words and phrases are always just a couple of mouse clicks away. This could help another researcher verify an analysis by reviewing the application of an important code or after regenerating an output report, and it helps the primary researcher to stay in close proximity to the source material even in the normally abstracted world of a survey.

Developing an analysis

In many survey situations the primary purpose of the analysis of the OEQ data may be to generate a table setting out the numbers or proportions of responses that have referred to each of the concepts identified in the data. This is a basic application of the inductive method. Providing that some important principles have been adhered to, such as restricting any thematic code to a single instance in any one response, the numerical count of occurrences for each code, which is a simple output of the CAQDAS program, will provide the required results. Inevitably some respondents will mention several concepts in their response to an OEQ and so these results will be similar to those obtained from a 'tick all that apply' closed question in a survey where a multiple response analysis results in more responses than cases appearing in frequency output tables. The point is not to mimic the analysis available from the main survey data by looking at the proportions of the OEQ responses relating to given concepts, but to get a sense of proportion that helps the analyst identify the dimensions of

attitudes expressed in the OEQs, and to characterise the qualities of the subset offering OEQ responses.

An advantage of OEQ data is that the analyst has access to the respondents' ideas in their own words. This opens up possibilities of using content analysis and text mining tools to search for more subtle patterns within the responses. For example, it may be that older respondents use different terms to those employed by younger people to describe some shared concept, which may lead to an unexpected insight. Or it may be possible to identify more dimensions within attitudes to an issue than are revealed through simple agree/disagree Likert scales. If a survey will be repeated several times, say to measure some change over time, then the effort taken to build concept dictionaries in tools like QDA Miner's WordStat module will pay increasing dividends with each subsequent wave of the survey, providing more reliable text mining results with each iteration. These approaches will produce valid results if the underlying data are sufficiently rich and accurate.

Threats to validity may come from the data collection process, for example when interviewers in a CAPI situation (whether face to face or telephone) are asked to type the responses as they hear them and are tempted to abbreviate or paraphrase, or when context effects are created in a questionnaire by placing an exploratory OEQ immediately after a series of closed questions on the same topic (so that many respondents feed back the terminology they have just heard or seen). It is also important to consider those that do not provide OEQs in response to a feeder question. In the interviewer-aided survey presented in this article, non-response to an open ended question seeking more information was very low and invariably recorded as 'DK' in the verbatim data. However, the intention here is not to replicate the analytic inferences that can be drawn from the quantitative data which take into

account non-responses in the calculation of valid percentages but to capture, where possible, the tone and nuance of the verbatim response, and here a non-response, recorded as ‘don’t know’, could be considered as valid data.

The spreadsheets CAQDAS packages generate with counts of selected codes applied to each case enable further statistical analysis. The results of an analysis in a CAQDAS program can generally be exported for further processing in a statistical package. Such a process may seem questionable to some qualitative researchers, but in the survey situation it is more likely that the necessary requirements of sampling method and size have been met, and so this may be a reasonable step. For instance, in the current survey, data could be exported to a statistical package to model whether the coded OEQ responses, in the form of binary variables (ie. presence/absence of coded word/phrase), in conjunction with other demographic variables, predicted whether the respondent felt their actions effective or not.

In purely qualitative projects, the essential features of CAQDAS packages are the support for coding, forming subsets of associated codes, and selective retrieval strategies. In the latter, a distinction is made between ‘single sort’ and ‘multiple sort’ retrievals. A single sort retrieves all data segments given a single code (eg. ‘advice = sandbag’) whereas a multiple sort retrieves only data segments assigned two or more codes that the analyst believes to be related (eg. ‘advice = sandbag’ AND ‘why sandbags not used = not necessary’).The resulting lists of data segments help identify the characteristics of cases or circumstances sharing a particular condition. A range of Boolean operators allow analytically-targeted retrievals under the conditions of set theory (eg. ‘If flooded = not flooded’ AND ‘advice = sandbag’ AND (‘why sandbags not used = none available’ OR ‘why sandbags not used = not given any’)).

CAQDAS is not the only software support for analysing OEQs. If users want to take a quantitative content analytic approach a range of statistical packages may be considered, alongside hybrid programs like QDA Miner with WordStat add-on that offer both content analysis and CAQDAS-type analytic features. A package that may be familiar to survey researchers is Text Analytics for Surveys (TAS), developed by IBM SPSS. This software claims to automate the coding of OEQs using natural language processing (NLP). The term 'text analytics' developed within this field is roughly synonymous with the older term 'text mining'. Following the import of data (in various formats including SPSS data files), TAS builds terms (either simple words or phrases) from which the codes or categories are developed. For those engaged in customer, employee or product satisfaction surveys, this coding process is greatly facilitated by pre-built sets of categories shipped with the software. However, for all other research purposes the process essentially requires manual categorization. Here, conditional rules (ie. using combinations of the original extracted terms and Boolean operators) can be developed and saved. Once saved, future coding of new data from repeat survey administrations is open to automation, and the extensive time invested in coding is then repaid. 'Categorisation' work in TAS is equivalent to the development of 'dictionaries' in WordStat. However, while both packages provide visualisation procedures, that in TAS is provided as an aid to the categorisation process while that in QDAMiner is additionally used as an analysis tool. Thus while categorisation within TAS results in qualitative data being reduced to limited sets of categories or codes, it is then necessary to export this data to other software for any further analysis.

Relatedly, Schmidt (2010) discusses the use of CATPAC (whose word association features employ an algorithm based on a neural network) in combination with the Clementine

extension of SPSS, which enables users to perform data mining and multivariate analysis; Schmidt highlights the use of rule-based webs and multiple correspondence analysis. Such content analysis techniques are good at identifying clusters of associated words or terms, and this can be useful where there is analytic interest in patterns of words in a text. Such analyses are systematic and precise, but relatively shallow compared with qualitative data analysis (in traditions like Grounded Theory or interpretive phenomenology). However, one can profitably use such content analysis to derive hypotheses for deeper investigation using qualitative data analysis and CAQDAS. Facilitators of such investigations include the Boolean operators permitting selective retrievals using AND, OR, XOR, NOT relations and additional features to visualize and present textual data.

Conclusions

There are many degrees of data integration, from the use of one type of data to illustrate another (e.g., using an interview extract to bring alive a survey finding) through to full-blown conversion of one type of data to another (the 'quantitization' of qualitative data), and each relies on a different warrant, bears distinctive threats to validity, and requires different means to respond to these threats. One relevant implication in the present context is the need for clarity about the kinds of support that qualitative software can provide when combining OEQ and fixed response data. Any transformations of the data must be justified by considerations like sample size, any systematic divergence between response from different methods must be squarely accounted for, and, where findings from different methods converge, it must be established that this is not an artifact of the characteristic biases of given methods. For most researchers combining OEQ and closed response survey data, the latter two considerations

are most pertinent, since full transformation of qualitative data by ‘quantitization’ is relatively unusual. Regarding convergence, while closed responses and OEQ are collected in the same administration of the instrument, the data are of different forms (scaled response, text) and thus bear different characteristic biases and pose different coding requirements. The principal contribution that qualitative software can make is therefore in supporting consistent coding (and enabling this to be checked using codebook tools and analytic memo features supporting code elaborations), providing data retrieval, sorting and categorizing tools for systematic data analysis, and offering features to visualize and present textual data.

Qualitative software has a part to play in assessing the reliability and validity of the conclusions that may be drawn from OEQ data or OEQ data in conjunction with closed-response data. It addresses this by allowing users quickly to go back to the context of extracts (including from data matrices or other tabulations) to inspect the changes made in data reduction and analysis. To address contradictory findings and deal with biases characteristic of qualitative data, such software provides instant re-assignment of codes across all data views so users can inspect the effects of different assignments of codes to the data (sometimes with such changes automatically revising abstract representations of the data, codes, and themes). Also an aid to resolving contradictions and tracking bias are qualitative software features promoting ‘transparency’, with automatically-compiled audit trails allowing users to trace the steps by which given codes were assigned, and annotation features allowing users to unravel why they have assigned a given code, compile survey-type codebooks if they wish, and emulate Glaser and Strauss’s (1967) practice of writing ‘analytic memoes’. Qualitative software can thus be an effective way to ‘open up’ open-ended survey data, with gains in reliability and validity that warrant their integration with fixed response

data.

References

Fielding, Jane, Kerrill Gray, Kate Burningham, and Diane Thrush. 2005. “Flood Warning for Vulnerable Groups: Secondary analysis of flood data.” *Environment Agency R & D Report W5C-018/2.* Bristol, UK, Environment Agency.

Fielding, Nigel. 2009. “Going out on a limb: postmodernism and multiple method research.” *Current Sociology*, 57:427-447.

Frisbie, Bruce, and Seymour Sudman. 1968. “The use of computers in coding free responses.” *Public Opinion Quarterly* 32:216-232.

Giorgetti, Daniela, and Fabrizio Sebastiani. 2003. “Automating survey coding by multiclass text categorization techniques.” *Journal of the American Society for Information Science and Technology* 54:1269-1277.

Glaser, Barney, and Anselm Strauss. 1967. *The Discovery of Grounded Theory*. Chicago: Aldine.

Higgins, S., K. Ford, and I. Oberski. 1996. “Computer-aided qualitative analysis of interview data.” *British Educational Research Conference*. Lancaster, UK.

Miles, Matthew B. 1979. “Qualitative Data as an Attractive Nuisance: The Problem of Analysis.” *Administrative Science Quarterly* 24:590-601.

Stable URL: http://www.jstor.org/stable/2392365

Montgomery, Andrew C., and Kathleen S. Crittenden. 1977. “Improving coding reliability for open-ended questions.” *Public Opinion Quarterly* 41:235-243.

Schmidt, Marcus. 2010. “Quantification of transcripts from depth interviews, open-ended responses and focus groups: Challenges, accomplishments, new applications and perspectives for market research.” *International Journal of Market Research* 52:483-509.

Web Links:

Analysing Survey Data using CAQDAS
http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/support/analysingsurvey/

CAQDAS Web site Working Papers
http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/resources/workingpapers/index.htm

ATLAS.ti - http://www.atlasti.com/

CAQDAS Networking Project - http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/

Key Survey - http://www.keysurvey.co.uk

MAXQDA - http://www.maxqda.com/

NVivo - http://www.qsrinternational.com/

Online QDA - http://onlineqda.hud.ac.uk/

QDA Miner - http://www.provalisresearch.com/

Text Analytics for Surveys - http://www.spss.com/software/statistics/text-analytics-for-surveys/

Survey Monkey - http://www.surveymonkey.com/

Survey Said - http://www.surveysaid.com/Survey_Said/Home.html

Table 1: Open-ended questions in the Post Events survey

Q37. Can you think of any better ways to warn people in the area when a flood is likely to happen?

Yes	1
No	2
Don't Know	

IF Q37=Yes , THEN ASK: Q38

Q38. What better ways can you think of?

PROBE FULLY RECORD VERBATIM

Q39. Were you at any stage before the flood given any ADVICE on what you should do to prepare for the flood?

Yes	1
No	2
Don't Know	

IF Q39 = Yes, THEN ASK: Q40

Q40. What advice were you given?

PROBE FULLY RECORD VERBATIM

Q41.This lists some of the types of advice which households MAY be offered in the event of a flood warning. Thinking about the recent flood, which of these, if any, do you remember being advised to do, whether or not you actually did them?

[.....]

If Q41 = 'Block doorways / airbricks with sandbags'

Q57. You said that you were advised to block doorways / airbricks with sandbags.

Did you manage to act on this advice at all?

Yes	1
No	2
Don't Know	

IF Q57 = No

THEN ASK: Q58

Q58. Why did you not manage to do this?

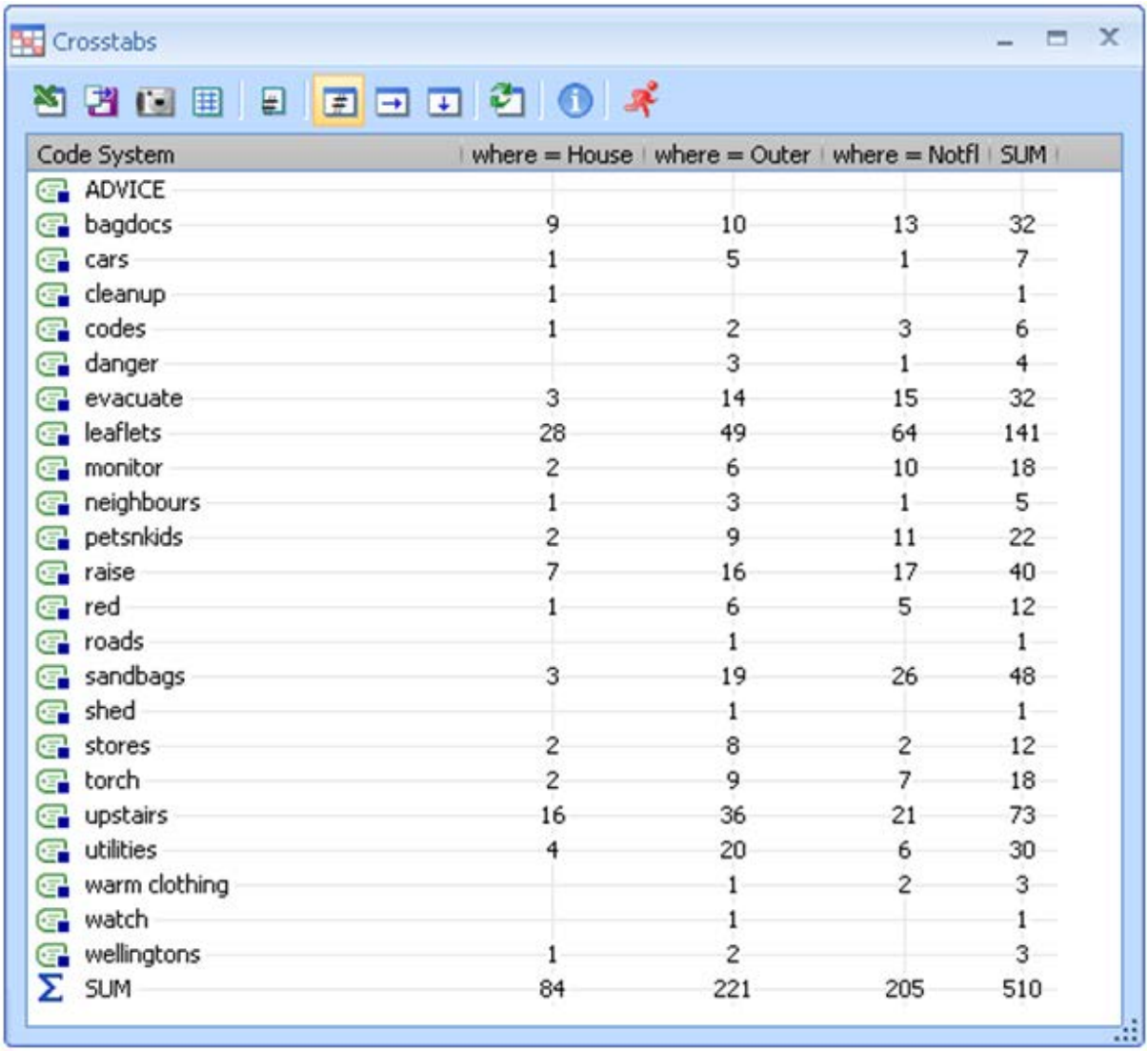
TYPE IN, RECORD VERBATIM

Table 2: Layout of the responses from the perspective of the variable/question

Variable Perspective	
Header: Question Name/ID - (e.g. Question 31)	
Case 1000	Response by Case 1000 to Q31
Case 1005	Response by Case 1005 to Q31
Case 1234	Response by Case 1234 to Q31
Case 1235	Response by Case 1235 to Q31
etc.	

Table 3: Layout of responses from the perspective of the case/respondent

Case Perspective	
Header: Respondent Name/ID - (e.g. Case 1234)	
Question 27	Response by Case 1234 to Q27
Question 31	Response by Case 1234 to Q31
Question 32	Response by Case 1234 to Q32
Question 44	Response by Case 1234 to Q44
etc.	



Code System	where = House	where = Outer	where = Notfl	SUM
ADVICE				
bagdocs	9	10	13	32
cars	1	5	1	7
cleanup	1			1
codes	1	2	3	6
danger		3	1	4
evacuate	3	14	15	32
leaflets	28	49	64	141
monitor	2	6	10	18
neighbours	1	3	1	5
petsnkids	2	9	11	22
raise	7	16	17	40
red	1	6	5	12
roads		1		1
sandbags	3	19	26	48
shed		1		1
stores	2	8	2	12
torch	2	9	7	18
upstairs	16	36	21	73
utilities	4	20	6	30
warm clothing		1	2	3
watch		1		1
wellingtons	1	2		3
SUM	84	221	205	510