This is the authors' final version of an article published in Psychology and Health 2010;25(10):1229-1245. The original article is available at: http://www.informaworld.com/smpp/content~db=all~content=a916180756~frm=titlelink?words=adequate|sample|size&hash=2779130764

What is an adequate sample size? Operationalising data saturation for theory-based interview studies

Abstract

In interview studies, sample size is often justified by interviewing participants until reaching "data saturation". However, there is no agreed method of establishing this. We propose principles for deciding saturation in theory-based interview studies (where conceptual categories are pre-established by existing theory). First, specify sample size for initial analysis (minimum analysis sample). Second, specify how many more interviews will be conducted without new ideas emerging (stopping criterion). We demonstrate these principles in two studies, based on Theory of Planned Behaviour, designed to identify three belief categories (Behavioural, Normative, Control), using a minimum analysis sample of 10 and stopping criterion of 3. Study 1 (retrospective analysis of existing data) identified 84 shared beliefs of 14 general medical practitioners about managing patients with sore throat without prescribing antibiotics. The criterion for saturation was achieved for Normative beliefs but not for other beliefs or for study-wise saturation. In Study 2 (prospective analysis), 17 relatives of people with Paget's disease of the bone reported 44 shared beliefs about taking genetic testing. Study-wise data saturation was achieved at interview 17. We propose specification of these principles for reporting data saturation in theory-based interview studies. The principles may be adaptable for other types of interview studies.

Keywords: data saturation; sample size; interviews as topic; models, psychological; theory-based content analysis.

What is an adequate sample size for interview studies? Operationalising data saturation for theory-based content analysis

Background

In studies that use semi-structured interviews that are analysed using content analysis, sample size is often justified on the basis of interviewing participants until "data saturation" is reached. However, there is no agreed method of establishing when data saturation has been reached and so it is not clear what this means in practice. In this paper we propose a method for establishing and reporting how data saturation has been achieved in theory-based interview studies (i.e., in which conceptual categories are pre-established from existing theory). We suggest a set of systematic principles by which researchers can report their justification for the decision that an appropriate sample size has been attained in such interview studies. In addition, we suggest how this might be tested.

The concept of data saturation was introduced to the field of qualitative research by Glaser and Strauss (1967) and referred to the point in data collection when no new additional data are found that develop aspects of a conceptual category. The idea of data saturation is a very useful guide for such research, in which the appropriate sample size is a function of the purpose of the study and the complexity, range and distribution of experiences or views of interest, rather than of the statistical parameters used in quantitative research (for example, in the form of a power analysis). Indeed, Guest, Bunce and Johnson (2006) claim that "saturation has … become the gold standard by which diversity samples are determined in health science research" (p. 60). In the context of interview studies where the conceptual categories, or constructs, are pre-established on the basis of existing theory, if sampling is adequate (and if the interviews have been

effective in eliciting participants' experiences or views within these conceptual categories), it is likely that the content domain of the construct has been adequately populated (or saturated). Data saturation is an important concept as it addresses whether such a theory-based interview study is likely to have achieved an adequate sample for *content validity*.

The question of sample size is also important because the use of samples that are larger than needed is an ethical issue (because they waste research funds and participant time) and the use of samples that are smaller than needed is both an ethical and a scientific issue (because it may not be informative to use samples so small that results reflect idiosyncratic data and are thus not transferable, and may therefore be a waste of research funds and participant time).

The idea of sampling until data saturation is achieved has been invoked in research for some time in several health-related disciplines. To get a sense of the way the term has recently been used in disciplines that focus on health research, we reviewed all papers published in the multidisciplinary journal *Social Science and Medicine* during the 16-month period June 2006 to September 2007 (inclusive). 'Data saturation' was mentioned in 18 papers, of which 15 claimed to have achieved data saturation. The definitions were consistent; data saturation meant that no new themes, findings, concepts or problems were evident in the data. However, it was not clear how data saturation was decided. Table 1 provides the relevant quotations from each of the studies reviewed, showing how saturation was defined and justified. This paper addresses the questions: What does data saturation mean in practice? As a research community, how might we agree principles so that research teams can decide when it has been reached? How can researchers best

present evidence to specify or defend the judgement that data saturation has been achieved in a way that is transparent to readers?

TABLE 1 HERE

The question addressed in this paper, then, is 'What does it mean, in practice, to say that *NO* new themes have emerged?' If a second participant is very similar to the first insofar as s/he does not mention any new ideas, it is clearly not appropriate to stop interviewing after two interviews. Yet, how many interviews with no new ideas does it take before the researcher may be confident that no more importantly new ideas would be mentioned if more participants were sampled? The question might need to be answered differently depending on the research question and type of interview study. Some forms of analysis (e.g., grounded theory) seek to build theory by identifying constructs implied by the data and building them into a network of associations, whereas in theory-based content analysis, the researcher seeks to use the data to populate pre-specified theoretical constructs with contextually relevant content. In this paper we focus on studies in which interviews are used to generate data to populate pre-specified theoretical constructs with contextually relevant content. We suggest some principles for deciding that data saturation has been reached and for reporting evidence of data saturation. We illustrate the proposed principles in two studies that stimulated our interesting in this topic. They involved theory-based content analysis of theoretically-focused interview transcripts founded on the Theory of Planned Behaviour (Ajzen, 1991). We acknowledge that this may not apply to other approaches to analysis.

The TPB provides a theoretical framework for predicting intentions and behaviour. It has received substantial empirical support from systematic reviews of correlational studies (e.g., Armitage & Conner, 2001) and experimental studies (e.g., Webb & Sheeran, 2006). TPB research uses standard methods (e.g., Francis, Eccles, Johnston, Walker, Grimshaw, Foy, et al, 2004) to operationalise the constructs in the model: Attitude (how much the person is in favour of performing a specified behaviour), Subjective Norm (how much the person feels pressure from social sources to perform the behaviour, or not), and Perceived Behavioural Control (how much the person feels that the behaviour is within his or her control). Each of these variables (Attitude, Subjective Norm and Perceived Behavioural Control, or PBC) is measured by asking participants to complete a questionnaire by reporting the extent to which they agree or disagree with items reflecting three kinds of specific beliefs. *Behavioural beliefs* (the perceived advantages and disadvantages of enacting the behaviour) are proposed determinants of Attitude. *Normative beliefs* (the individuals or social group perceived to exert pressure to enact the behaviour, or not) are proposed determinants of Subjective Norm. *Control beliefs* (the perceived factors that make it easier or more difficult to enact the behaviour) are proposed determinants of PBC.

Rather than using a 'one-size-fits-all' questionnaire, the TPB stipulates that the questionnaire items should reflect issues that are relevant to the target behaviour for the population to be investigated. Existing guidance on conducting these interviews does not specify the number of interviews necessary. Interview transcripts are subjected to theory-based content analysis and Ajzen (1988) has provided detailed guidance on the interview format. The objective of the analysis is to discover, from interviewees, what are most

'salient' Behavioural, Normative and Control beliefs. This is done by identifying the views or beliefs that are most frequently mentioned, independently, by participants, in response to open questions. For this reason, the studies reported here analysed data saturation for *shared* beliefs (i.e., mentioned by two or more participants), as idiosyncratic beliefs (i.e., mentioned by only one participant) were not likely to be relevant to most of the population from which the participants were drawn.

This theory-based approach thus differs importantly from other types of qualitative research. First, in some studies, themes that appear to be 'idiosyncratic' within an initial sample might lead to further sampling of participants from potentially under-represented sub-groups for whom such themes might be important. Second, some studies explicitly search for contrasts within the sample in order to generate hypotheses about how individuals or sub-groups might differ. The principles for establishing data saturation that are proposed here do not apply to these other types of research. We suggest, however, that the principles might be adaptable to these kinds of studies, because the question, when to stop sampling, may significantly influence research findings and therefore may require team decisions that have a clear justification. As indicated above, we first propose the principles within the less complex context of an interview study based on pre-specified theoretical constructs. As long as an appropriately diverse sample has been used, the principles may justify a claim that data saturation has been achieved.

Principles for specifying data saturation

We propose four principles for analysis and reporting. First, researchers should specify *a priori* at what sample size the first round of analysis will be completed (in order to identify a basis for progressive judgements about data saturation). We will refer to this

as the *minimum analysis sample*. The specific number will depend on the complexity of the research questions and of the interview topic guide, the diversity of the sample and the nature of the analysis (e.g., the number and likely dimensionality of the target constructs). Of course, sampling would be conducted according to pre-specified 'stratification' factors that are relevant to the study (e.g., age, gender, rurality, ethnicity). Otherwise, spurious early data saturation may be achieved due to spurious homogeneity of the sample. (If many stratification factors are likely to be relevant to the research questions, a larger initial analysis sample is likely to be needed.)

The second principle is that researchers should specify *a priori* how many *more* interviews will be conducted, without new shared themes or ideas emerging, before the research team can conclude that data saturation has been achieved. We will refer to this as the *stopping criterion*. Analysis then proceeds on an ongoing basis until the stopping criterion is met.

To illustrate these two principles in the studies reported in the current paper, we specify the first two principles as follows (assuming two or three main stratification factors):

- Initial analysis sample: At least 10 interviews will be conducted (with appropriate diversity sampling).
- Stopping criterion: After 10 interviews, when three further interviews have been conducted with no new themes emerging, we will define this as the point of data saturation. The stopping criterion is tested after each successive interview (i.e., 11, 12 and 13; then 12, 13 and 14, and so on) until there are three consecutive interviews without additional material. In this phase of the study a research team might decide to specify

other groups of participants to sample, if analysis suggests that the stratification factors applied for the initial analysis sample may be inadequate.

In the interests of providing further deliberation, we offer additional principles. The third principle is that the analysis would ideally be conducted by at least two independent coders and agreement levels reported to establish that the analysis is robust and reliable. The fourth principle is that data saturation methods and findings ideally would be reported so that readers can evaluate the evidence. *A priori* criteria could be part of a paper's Methods section. We will demonstrate these principles below.

An earlier attempt has been made to specify a sample size rule for interview studies that are not theory-based. Guest et al. (2006) conducted interviews in two African countries on the topic of social desirability behaviour and accuracy of self-reported sexual behaviour. They documented the progression of theme identification after successive sets of six interviews, until 60 interviews had been conducted. Ninety-two percent of all codes were identified after 12 interviews and 97% of the 'important' codes (operationalised as the number of individuals expressing the same idea) were identified within these 12 interviews. Guest et al. concluded that about 12 is a sufficient sample for interview studies analysed for emergent themes. However, they questioned the transferability of their findings. Furthermore, there appeared to be no 'development' of the interview process; the topic guide did not evolve to explore emerging themes in greater depth during the course of the interview study. In that sense, the methods used by Guest and colleagues were more like the pre-determined, theory-based approach described in the studies reported here than like an emergent themes analysis. In addition, as the analysis proceeded in sets of six, it is not clear when their identified level of

saturation was reached; it was somewhere between seven and 12. In contrast to this approach, we propose a set of principles for establishing the appropriate sample size, together with ways to present data to support this judgement.

In this paper we report data from two studies to illustrate and critically examine these principles. The data are presented in the form of cumulative frequency distributions, showing which participants mentioned a 'new' idea (or belief or theme), that is, a belief not previously elicited. As the objective of each of the studies was to elicit beliefs relating to three theoretical constructs (Attitude, Subjective Norm and PBC), we test the criterion for data saturation both at the level of each individual construct and at the study-wise level.

Study 1. Content analysis of general medical practitioners' beliefs about managing upper respiratory tract infections.

Background

This study used the Theory of Planned Behaviour to predict general medical practitioners' (GPs') intentions and behaviour relating to managing patients with upper respiratory tract infections (URTIs) without prescribing antibiotics. UK-based clinical guidelines recommend that GPs manage patient with URTI *without* prescribing antibiotics. The data reported here relate to a retrospective re-analysis of the first phase of a larger study that is reported elsewhere (Eccles, Grimshaw, Johnston, Steen, Pitts, Steen, et al, 2007).

This study of the clinical behaviour of healthcare professionals used the Theory of Planned Behaviour (TPB; Ajzen, 1991) as a theoretical framework for predicting intentions and actual prescribing behaviour. For the purpose of constructing a

questionnaire to measure three of the TPB constructs (Attitude, Subjective Norm and Perceived Behavioural Control), GPs were interviewed to identify their beliefs relating to managing patients with URTI without prescribing antibiotics. Interviews were conducted to elicit three kinds of beliefs: Behavioural beliefs, Normative beliefs and Control beliefs and the study team made the judgment that data saturation had been achieved. The objective of reporting the interview data here is to examine the sampling strategy used for this study and to find out (retrospectively) whether this judgment was consistent with the proposed principles for establishing data saturation.

Methods

Participants. Participants were 14 GPs (years in practice 4.5-25; 2 female; practising for 5-10 half-day sessions per week; from a range of regions in Scotland and north-east England). At the time, the research team felt that the first 10 participants (initial analysis sample) represented adequate diversity on these pre-specified stratification factors.

Materials. The interview topic guide was based on standard methods used for the TPB (Francis et al., 2004), i.e., questions about the advantages and disadvantages of managing patients with URTI without prescribing antibiotics, who might approve or disapprove of this behaviour; and what factors might make it easier or more difficult to do this.

Procedure. Semi-structured interviews, lasting approximately 40 minutes, were conducted with individual participants. The interviews were audiorecorded, transcribed, anonymised and content analysed.

Analysis. Theory-based content analysis was conducted in three steps. First, one researcher split each transcript into separate utterances. Second, one researcher grouped the utterances of different participants into similar beliefs and used wording from the

transcripts to describe each belief ("summary data"). Third, two judges independently coded each belief for the presence/absence of three kinds of belief: Behavioural belief, Normative belief and Control belief. Krippendorff's alpha (Krippendorff, 2004) was used to describe agreement between judges at the third step, separately for each construct.

Data saturation analysis was conducted in four steps. First, data tables were constructed at the level of specific beliefs elicited for each individual. Second, summary tables were constructed for each of the three kinds of belief to display the beliefs that were mentioned by each participant interviewed. This summary table contained binary (yes/no) data presented sequentially and included idiosyncratic beliefs (i.e., beliefs that were not shared by at least two participants). (See Appendix for format of table.) Third, data from the summary tables were used to construct a series of cumulative frequency graphs, one for each type of belief (Behavioural, Normative and Control) and one line for 'All beliefs'. These lines displayed, sequentially, the frequency with which each (shared) individual belief was mentioned by the 14 participants.

These cumulative frequency graphs were inspected to investigate: (a) the number of shared beliefs elicited by the initial analysis sample (which was set at 10); (b) the number of interviews required to meet the stopping criterion (which was set at three) for each construct, and overall; (c) whether any new shared beliefs emerged following three successive interviews with no new shared beliefs (for each construct and overall).

Results

Inter-rater reliability. Eighty-four summary beliefs (both shared and idiosyncratic) were identified and independently coded (by JF and CR) for presence/absence of Behavioural, Normative and Control beliefs. Krippendorff's alpha reliability estimates (1000 bootstrap

samples) were 0.67 for Behavioural beliefs, 0.93 for Normative beliefs and 0.63 for Control beliefs.

Summary data. Figure 1 presents cumulative frequency graphs for Participants 1 to 14, for the specific, shared beliefs about managing patients with URTI without prescribing antibiotics. The number sequence, "1 2 3" above a line highlights the application of the stopping criterion.

FIGURE 1 HERE

Construct-level saturation. From Figure 1, the line representing Behavioural beliefs shows that, when asked about advantages and disadvantages of managing patients with URTI without prescribing antibiotics, the first participant mentioned 19 distinct beliefs. After the fifth interview, 35 shared beliefs (i.e., beliefs mentioned by at least two participants) had been elicited.

After 10 interviews, the initial analysis sample yielded 36 shared Behavioural beliefs and there had been no new shared beliefs for two interviews. The following two interviews (11, 12) did not generate new shared beliefs but there was one new shared belief at interview 13. So applying the stopping criterion for construct saturation (i.e., three interviews with no new shared beliefs) indicates that saturation was not achieved. No new shared Behavioural beliefs were elicited at Interview 14 and there were 37 in total.

From Figure 1, the line representing Normative beliefs shows that, after 10 interviews, the initial analysis sample had yielded 11 shared beliefs. There was one new

shared belief in Interview 11 but no new shared beliefs in Interviews 12, 13 or 14. Application of the stopping criterion thus suggests that construct saturation was reached after 14 interviews. The line representing Control beliefs shows that, after 10 interviews, the initial analysis sample had yielded nine shared beliefs and there were no new shared beliefs in interviews 11 or 12 but there was one new shared belief at interview 13. So applying the stopping criterion for construct saturation indicates that saturation was not achieved. No new shared Control beliefs were elicited at Interview 14 and there were 10 in total.

Study-wise saturation: all belief categories. Finally, the line representing all belief categories in Figure 1 shows that, after 10 interviews, the initial analysis sample had yielded 57 shared beliefs and there were no new shared beliefs in interviews 11 or 12. However, there were two new shared beliefs at interview 13. So applying the stopping criterion indicates that study-wise saturation was not achieved, despite the research team's sense that data saturation had occurred.

However, this study was conducted before the proposed principles for establishing data saturation were devised. Fourteen interviews were conducted but two more interviews without new shared beliefs emerging would have been necessary to meet the proposed criterion for saturation. This is considered further in the General Discussion below.

Study 2. Content analysis of beliefs about genetic screening for Paget's disease of the bone

Background

This study investigated the acceptability of a potential genetic screening service for relatives of people with Paget's disease of the bone (PDB; Langston, Johnston, Robertson, Campbell, Entwistle, Marteau, et al., 2006) using the Theory of Planned Behaviour. The aim of the interview study was to identify the beliefs of a sample of individuals who were genetic relatives of people affected by PDB, with respect to a specific behaviour, taking a genetic test. Following TPB methodology, these beliefs were then used to generate questionnaire items for a subsequent study.

Study 1 had identified that, despite the research team's belief that saturation was achieved after 14 interviews, the 10+3 criterion for study-wise data saturation had not been met at that point. Study 2 used the stopping criterion for data saturation to decide the sample size in a contrasting sample.

Methods

Participants. Participants were 17 blood relatives (65% female) of people with a confirmed diagnosis of PDB. Of these, 76% were first degree relatives (11 children and two siblings) and 24% were second degree relatives (3 grandchildren and 1 first cousin). The research team felt that the first 10 participants (the initial analysis sample) represented adequate variation on these pre-specified stratification factors. (However, socio-economic categories were not recorded.)

Materials, Procedure Analysis. These replicated the methods of Study 1, except for one detail. In an attempt to improve the inter-rater reliabilities for Behavioural beliefs and

Control beliefs, an explicit decision rule was applied by raters. Control beliefs were defined as antecedents, i.e., factors that might occur before the behaviour was performed (e.g., *If I have no transport it will be more difficult for me to attend for a screening test*). In contrast, Behavioural beliefs were defined as consequences, i.e., factors that might occur after the behaviour was performed (e.g., *If I attend for a screening test I might worry about the result*).

Results

Inter-rater reliability. Forty-four summary beliefs (both shared and idiosyncratic) were identified and independently coded (by JF and CR), as for Study 1. Krippendorff's alpha reliability estimates (1000 bootstrap samples) were 0.85 for Behavioural beliefs, 1.00 for Normative beliefs and 0.86 for Control beliefs, indicating an improvement in inter-rater reliability with the application of the decision rule for distinguishing between Behavioural beliefs and Control beliefs.

Summary data. Figure 2 presents cumulative frequency graphs for Participants 1 to 17, for the specific, shared Behavioural, Normative and Control beliefs, and all beliefs about attending a screening test for PDB. Again, the number sequence, "1 2 3" above or below a line highlights the application of the stopping criterion.

FIGURE 2 HERE

Construct-level saturation. Figure 2 shows that, when asked about advantages and disadvantages of taking a screening test for Paget's disease, the first participant

mentioned four distinct behavioural beliefs. After the fourth interview, 11 shared Behavioural beliefs had been elicited.

After 10 interviews, the initial analysis sample yielded 12 shared Behavioural beliefs. However, new shared beliefs were elicited in interviews 11 and 12. There were no new shared beliefs in interviews 13, 14 or 15, so applying the stopping criterion for construct saturation (i.e., three interviews with no new shared beliefs) indicates that saturation was achieved after 15 interviews. If sampling had ceased at this point, no shared Behavioural beliefs would have been missed compared with the data provided by the full sample of 17. In all, 14 shared behavioural beliefs were elicited.

From Figure 2, after 10 interviews, the initial analysis sample had yielded eight shared Normative beliefs. One additional belief was elicited at interview 11. Application of the stopping criterion thus suggests that construct saturation was reached after 14 interviews and, if interviewing had ceased at that point, no further Normative beliefs would have been missed (within the sample of 17). After 10 interviews, the initial analysis sample had yielded 11 shared Control beliefs. The criteria for saturation were met after 13 interviews, but if interviewing had ceased at that point, one further Control belief would have been missed (at interview 14).

Study-wise saturation: all belief categories. Finally, from Figure 2, the line representing all belief categories shows that, after 10 interviews, the initial analysis sample had yielded 31 shared beliefs. Interviews 11 and 12 generated three new shared beliefs. In interview 13 there were no new beliefs but one further belief was elicited in interview 14. Study-wise data saturation was achieved after 17 interviews, and so interviewing ceased at that point. The total number of shared beliefs elicited in the study was 35.

General Discussion

Specifying the principles of data saturation (purposive diversity sampling for a minimum of 10 interviews, three further interviews with no new themes and presentation of data sequentially as cumulative frequency graphs) enabled the Study 2 research team to agree, and report, the point at which data saturation was achieved, in a transparent and reliable manner (assuming appropriate conduct of the interviews and reliability of coding). By distinguishing between construct saturation and study-wise saturation it was possible to assess saturation and adequacy of sampling at different levels. This contrasts with Study 1, in which the study team had judged, subjectively, that saturation was achieved but retrospective application of the stopping criterion suggested that at least two more interviews would be necessary to demonstrate that the criterion had been met.

Is the proposed criterion too stringent? Inspection of the results from Study 2 at the construct level may help to answer this question. If the stopping criterion had been applied at the construct level, then a study to investigate only Behavioural beliefs would have ceased sampling after 15 interviews; a study to investigate only Normative beliefs would have ceased sampling after 13 interviews; and a study to investigate only Control beliefs would have ceased sampling after 14 interviews. If this had occurred, one shared belief from the sample interviewed (out of the total of 35), or 3%, would have been missed. This is consistent with the findings of Guest and colleagues, who reported that the first 12 interviews elicited 97% of the important codes out of a total of 60 interviews. Thus, although the 10+3 criterion is not perfect, it appears to be a fairly effective guide (in the same way that the 0.05 significance criterion for quantitative studies allows that a Type 1 error may be made in approximately 5% of studies). We therefore suggest that this approach has proved robust for these examples of theory-based analysis. The

principles may be adaptable for using and testing in further studies based on different theoretical assumptions or addressing different kinds of research questions. Such parallels with the principles of quantitative research may strengthen some interview studies but may not be applicable to all research paradigms.

While the 10+3 criterion should be tested further, we suggest that some accepted convention for agreeing data saturation could be helpful. Like the 0.05 significance criterion for quantitative studies, such a convention would be somewhat arbitrary and may not be helpful for researchers who disagree with attempts to appraise qualitative research according to fixed criteria. Other researchers might find it a useful point of reference for deciding when it was necessary to deviate from the convention where the objectives of the study required a more, or less, stringent criterion.

Some similarities and differences between the two studies reported here illustrate some further benefits of graphical presentation of the results of theory-based content analysis. One clear difference between the studies is the difference in the number of behavioural beliefs elicited. It appears that the advantages and disadvantages of performing these two behaviours, or the way the participants think about them, have different levels of cognitive complexity. One practical implication of this in the current context is that it would require many more questionnaire items to achieve content validity for the behaviour, managing patients with URTI without prescribing antibiotics, than for the behaviour, taking a genetic test for Paget's disease. Yet, it is noteworthy that a similar sample size for the two studies generated these large differences in cognitive complexity of the content.

A similarity between the studies is that, despite contrasting types of behaviour and people sampled, the number of new beliefs elicited started to plateau after around six interviews (although we would not claim that saturation was reached at this point as the first six interviews generated only 92% and 86% of shared beliefs in Studies 1 and 2, respectively). It is likely that the use of purposive diversity sampling for the first 10 interviews contributed to achieving this plateau so early. This permits some confidence that setting the minimum sample size at 13 is very likely to capture almost all the beliefs relating to Attitude, Subjective Norm and Perceived Behavioural Control. Presentation of the data in a similar way for content analysis based on other theories could similarly help assess which minimum analysis sample size and stopping criterion are appropriate.

There are of course several limitations to the principles proposed here. First, the actual numbers proposed for the minimum analysis sample and stopping criterion would require a body of evidence to demonstrate their appropriateness. Furthermore, it is possible that the appropriateness of particular conventions might vary across studies with different objectives and using different theoretical constructs (but this is clearly testable). It is the principle of specifying a minimum number of interviews and then a further number that generate no new ideas that we propose may be an important tool for specifying saturation. Second, the principles rely on high quality data collection. That is, appropriately trained and skilled interviewers who are able to use prompts, reflection and encouragement to elicit participants' views without asking leading questions or pre-empting interpretations are an essential part of the research process. Third, the analyses reported here assume clarity among the coders about what constitutes a single belief. This assumption appeared to be non-problematic in the special case of analysis based on the

TPB but such judgements may not be as clear in other types of studies. This would be important in distinguishing between idiosyncratic and shared beliefs. For example, in a study investigating individuals' perceived consequences of taking a screening test, the ideas *"I might get twitchy about the results"* and *"I might get anxious about the results"* might be regarded as the same belief. However, in a study to investigate the kinds of words that individuals use to describe their emotions about screening, these two utterances could demonstrate important differences.

Fourth, these criteria have been applied only to the particular type of research involving content analysis based on the Theory of Planned Behaviour. Whether the principles of data saturation that we propose are appropriate for application to other types of interview study (e.g., those using other theories, or grounded theory; Bryant & Charmaz, 2007) would require further investigation. In particular, other types of research may focus on the elicitation of novel ideas that would then be pursued further with carefully sampled participants, or on contrasts and contradictions within and between participants. These types of research questions may require sample sizes that differ markedly from the sizes proposed here. However, we suggest that the basic ideas of specifying an initial analysis sample and developing some kind of stopping criterion may be helpful in deciding an appropriate sample size in the context of other types of research (for example, to help think about appropriate sub-samples to address sub-questions).

Finally, we acknowledge that practical constraints involving research staff or timelines may make it not always possible to apply the proposed principles. Practical issues may restrict a research team's ability to conduct ongoing analysis of interviews or to present the data in the ways we have illustrated. However, the principle of monitoring

additional material that emerges in consecutive interviews may be helpful in managing the research process.

In conclusion, we offer the following recommendations for future interview studies that use theory-based content analysis. First, researchers could specify *a priori* their criteria for study-wise data saturation in study protocols (deciding the size of the initial analysis sample and the stopping criterion) and report these criteria in publications (including publications of protocols). Second, data could effectively be organised and presented using cumulative frequency graphs, as illustrated here, to enhance the transparency and verifiability of the decision that saturation is achieved and to address different kinds of research topics (such as descriptions of the complexity or multifaceted nature of certain issues for certain participant groups). Third, a body of evidence could thereby be accumulated to establish a convention for decisions about sample sizes in different types of interview study. There is a need for further research to reflect on and develop this idea.

Acknowledgements. The PRIME study (Study 1) and the GaP study (Study 2) were funded by the UK Medical Research Council. We thank the participants in both studies for generously sharing their views. Jeremy Grimshaw holds a Canada Research Chair in Health Knowledge Transfer and Uptake. Jill Francis is funded by the Chief Scientist Office of the Scottish Government Health Directorates.

References (* denotes references included in the reported review (Table 1)).

Ajzen, I. (1988). *Attitudes, personality and behaviour*. Milton Keynes; OUP.

Ajzen, I. (1991). The theory of planned behaviour. *Organizational Behaviour and Human Decision Processes*, 50, 179-211.

Armitage, C.J., & Conner, M. (2001). Efficacy of the Theory of Planned Behaviour: A meta-analytic review. *British Journal of Social Psychology*, 40, 471-499.

Bryant, A., & Charmaz, K. (eds), *The SAGE handbook of grounded theory*. Sage Publications; London.

*Bryant, J., Porter, M., Tracy, S.K., & Sullivan, E.A. (2007) Caesarean birth: Consumption, safety, order, and good mothering. *Social Science & Medicine*, 65, 1192-1201.

*Bugge, C., Entwistle, V.A., & Watt, I.S. (2006). The significance for decision-making of information that is not exchanged by patients and health professionals during consultations. *Social Science & Medicine*, 63, 2065-2078.

*Caldwell, P.H., Arthur, H.M., Natarajan, M., & Anand, S.S. (2007). Fears and beliefs of patients regarding cardiac catheterization. *Social Science & Medicine* 65, 1038-1048.

*Damschroder, L.J., Pritts, J.L., Neblo, M.A., Kalarickal, R.J., Creswell, J.W., & Hayward, R.A. (2007). Patients, privacy and trust: Patients' willingness to allow researchers to access their medical records. *Social Science & Medicine*, 64, 223-235.

*Devine, C.M., Jastran, M., Jabs, J., Wethington, E., Farell, T.J., & Bisogni, C.A. (2006). "A lot of sacrifices:" Work–family spillover and the food choice coping strategies of low-wage employed parents. *Social Science & Medicine*, 63, 2591-2603.

Eccles, M.P., Johnston, M., Hrisos, S., Francis, J., Grimshaw, J., Steen, N., Kaner, E.F. (2007). Translating clinicians' beliefs into implementation interventions (TRACII): A protocol for an intervention modeling experiment to change clinicians' intentions to implement evidence-based practice *Implementation Science*, 2, 27.

Francis, J.J., Eccles, M.P., Johnston, M., Walker, A.E., Grimshaw, J.M., Foy, R., et al. (2004). *Constructing questionnaires based on the theory of planned behaviour. A manual for health services researchers.* Centre for Health Services Research, University of Newcastle upon Tyne, UK. ISBN 0-9540161-5-7 and www.rebeqi.org.

Glaser, B.G., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago, IL: Aldine.

Godin, G., & Kok, G. (1996). The Theory of Planned behaviour: A review of its applications to health-related behaviours. *American Journal of Health Promotion, 11*, 87-98.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, 18, 59-82.

*Hawkins, R.L., & Abrams, C. (2007). Disappearing acts: The social networks of formerly homeless individuals with co-occurring disorders. *Social Science & Medicine*, 65, 2031-2042.

*Hsiao, A.F., Ryan, G.W., Hays, R.D., Coulter, I.D., Andersen, R.M., & Wenger, N.S. (2006). Variations in provider conceptions of integrative medicine. *Social Science & Medicine*, 62, 2973-2987.

Krippendorff. K. (2004). *Content Analysis: An Introduction to Its Methodology* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Langston, A.L., Johnston, M., Robertson, C., Campbell, M.K**.,** Vikki A Entwistle, V.A., Marteau, T.M., McCallum, M., & Ralston, S. H. (2006). Protocol for stage 1 of the GaP study (Genetic testing acceptability for Paget's disease of bone): an interview study about genetic testing and preventive treatment: would relatives of people with Paget's disease want testing and treatment if they were available? *BMC Health Services Research*, 6, 71.

*Leavey, G., Loewenthal, K., & King, M. (2007). Challenges to sanctuary: The clergy as a resource for mental health care in the community. *Social Science & Medicine*, 65, 548-559.

*Lonardi, C. (2007). The passing dilemma in socially invisible diseases: Narratives on chronic headache. *Social Science & Medicine*, 65, 1619-1629.

*Murray, J., Banerjee, S., Byng, R., Tylee, A., Bhugra, D., & Macdonald, A. (2006).Primary care professionals' perceptions of depression in older people: a qualitative study. *Social Science & Medicine*, 63, 1363-1373.

*Poletti, T., Balabanova, D., Ghazaryan, O., Kocharyan, H., Hakobyan, M., Arakelyan, K., & Normand, C. (2007).The desirability and feasibility of scaling up community health insurance in low-income settings—Lessons from Armenia. *Social Science & Medicine*, 64, 509-520.

*Rapley, T., May, C., & Kaner, E.F. (2006). Still a difficult business? Negotiating alcohol-related problems in general practice consultations. *Social Science & Medicine*, 63, 2418-2428.

*Ungar, W.J., Mirabelli, C., Cousins, M., & Boydell, K.M. (2006). A qualitative analysis of a dyad approach to health-related quality of life measurement in children with asthma. *Social Science & Medicine*, 63, 2354-2366.

*Wainberg, M.L., González, M.A., McKinnon, K., Elkington, K.S., Pinto, D., Gruber Mann, C., & Mattos, P.E. (2007). Targeted ethnography as a critical step to inform cultural adaptations of HIV prevention interventions for adults with severe mental illness. *Social Science & Medicine*, 65, 296-308.

*Wanich Bradway, C., & Barg, F. (2006). Developing a cultural model for long-term female urinary incontinence. *Social Science & Medicine*, 63, 3150-3161.

Webb, T. L., & Sheeran, P. (2006). Does changing behavioural intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin*, 132, 249-268.

*Webber, M.P., & Huxley, P.J. (2007). Measuring access to social capital: The validity and reliability of the Resource Generator-UK and its association with common mental disorder. *Social Science & Medicine*, 65, 481-492.

*Williams, B., Mukhopadhyay, S., Dowell, J., & Coyle, J. (2007). From child to adult: An exploration of shifting family roles and responsibilities in managing physiotherapy for cystic fibrosis. *Social Science & Medicine* 65, 2135-2146.

*Wray, N., Markovic, M., & Manderson, L. (2007). Discourses of normality and difference: Responses to diagnosis and treatment of gynaecological cancer of

Australian women. *Social Science & Medicine*, 64, 2260-2271.

Figure 1. Cumulative frequency of shared behavioural beliefs, normative beliefs, control beliefs, and all beliefs elicited by interviews for the behaviour, ‘managing patients with upper respiratory tract infection without prescribing antibiotics’.

Figure 2. Cumulative frequency of shared behavioural beliefs, normative beliefs, control beliefs, and all beliefs elicited by interviews for the behaviour, ‘taking a genetic test’.

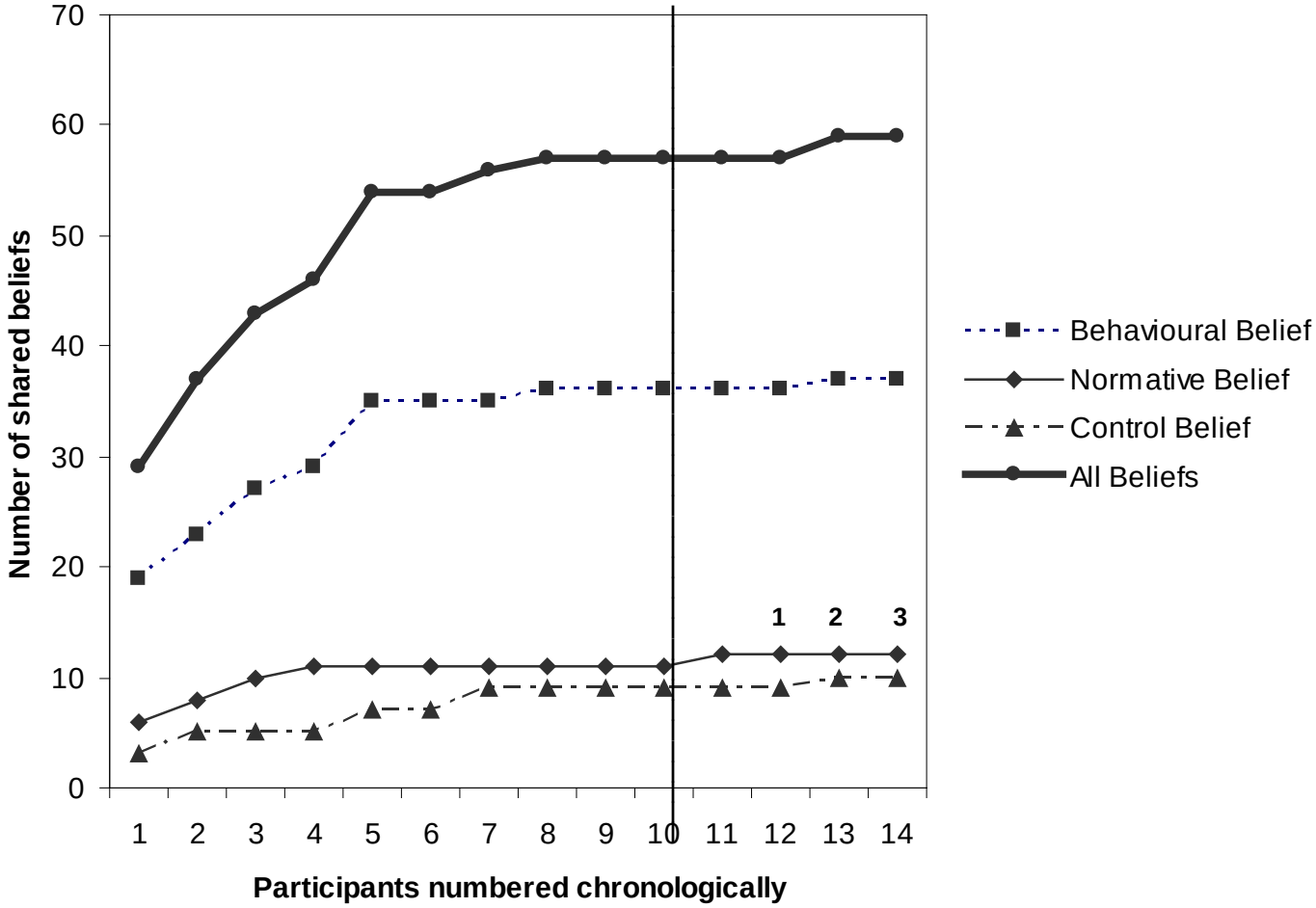
Table 1. Papers published in *Social Science and Medicine* containing reference to 'data saturation' (16-month period June 2006 – September 2007).

	Reference	**Quote**
1	Bryant, Porter, Tracy et al., 2007	Data were analysed for thematic content. Recruitment of participants ended once thematic saturation was achieved or no new themes were observed in the data (Guest, Bunce, & Johnson, 2006). p.1194
2	Caldwell, Arthur, Natarajan et al., 2007	Recruitment continued until saturation of themes was determined to have occurred (Strauss & Corbin, 1998, p1041). p.1041
3	Williams, Mukhopadhyay, Dowell et al., 2007	Sampling continued until ongoing analysis revealed no new findings, and saturation was obtained (Strauss & Corbin, 1990) (Tables 1 and 2). p.2137
4	Lonardi, 2007	Recruitment ceased as the main concepts started to show redundancy along the various stories, and trajectory models started to show clear shapes according to the criterion of theoretical saturation (Glaser & Strauss, 1967). p.1621
5	Hawkins & Abrams, 2007 [a]	Of the 39 participants, six did not complete a second interview because they were unavailable, impaired, or the research team felt the first interview had achieved saturation. p.2035
6	Webber & Huxley, 2007	The sample size was determined by the principle of theoretical saturation (Coyne, 1997), so we continued the interviews until no new problems emerged with the instrument. p.484
7	Leavey, Loewenthal & King, 2007	This is a qualitative study using a purposive sampling or theoretical strategy generally associated with grounded theory whereby the collection and analysis of the data are inter-related (Strauss & Corbin, 1990, p. 67). Thus, the data gathered from the preliminary interviews informed the direction of further data collection and informant selection. This helped in the exploration of the parameters of the study and provided opportunities for increasing the 'density' and 'saturation' of significant, recurring and ambiguous categories. p.549
8	Wainberg, González, McKinnon, et al., 2007	Saturation (i.e., a sufficient number of field observations) was reached for specific time periods of the institutions' operating hours and on specific days in accordance with principles of grounded theory in qualitative research (Strauss & Corbin, 1994). p.298
9	Wray, Markovic & Manderson, 2007	This interviewing technique ensured that study participants spoke about issues pertinent to their experience of illness and helped achieve data saturation. p.2263 Data collection and data analysis were conducted concurrently, allowing us to modify interview guidelines for subsequent in-depth interviews to incorporate new emerging themes and to stop further recruitment on achieving data saturation. p.2263
10	Poletti, Balabanova, Ghazaryan,	Following a grounded theory approach, topics were covered in the interviews until saturation was

	et al., 2007	reached (Strauss & Corbin, 1998); subsequent interviews focused on filling gaps in the data.[a] p.511
11	Damschroder, Pritts, Neblo, et al., 2007	We achieved theme saturation before reaching the end of the sample of 16 coded transcripts, which also contributes to coding trustworthiness (Miles & Huberman, 1994). p.226
12	Bradway & Barg, 2007	Theoretical saturation was achieved after interviews with 17 women were completed. p.3152
13	Devine, Jastran, Jabs, et al., 2006	We continued recruiting and analysis until no new themes emerged from interviews, and theoretical saturation was reached (Sobal, 2001). p.2594
14	Ungar, Mirabelli, Cousins, et al., 2006	Saturation was achieved at 16 dyad interviews. p.2354 Enrollment continued until saturation occurred, which was achieved at 16 dyads. p.2355 Data collection and open coding continued until new information produced little or no change to data categories, i.e. until theoretical saturation was achieved (Sandelowski, 1995). This occurred when the 16th dyad interview was completed and analysed. p.2356
15	Rapley, May & Kaner, 2006	The analysis developed until category saturation was reached (i.e. interviews and analytic procedures yielded no new material for analysis). The analysis was further developed and validated through the group interviews. p.2420
16	Bugge, Entwistle & Watt, 2006	Also, we did not sample to theoretical saturation, so it is possible that there are types of information that are sometimes not exchanged— either in the settings that we have studied or in other contexts— that we have missed. p.2074
17	Murray, Banerjee, Byng, et al., 2006	Data collection proceeded concurrently with analysis until theoretical saturation was achieved, in other words, until no new themes were emerging. p.1365
18	Hsiao, Ryan, Hays, et al., 2006	Theoretical saturation occurred after 50 interviews. p. 2975

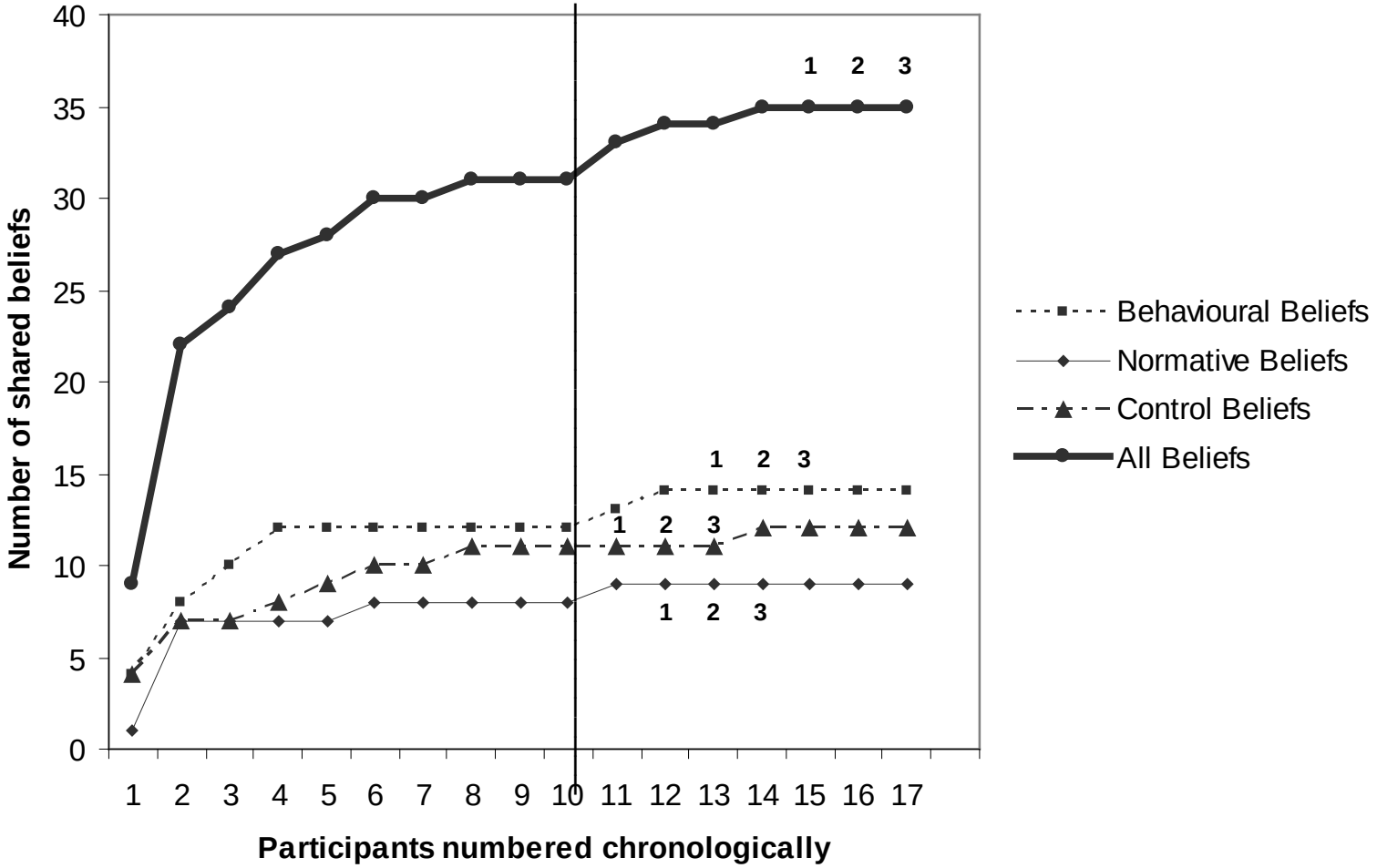
[a] This paper appeared to use this term as a within-participant term, i.e., the interviewer felt that it was appropriate to end the interview because the interviewee had said all that s/he wanted to say about the topic. This type of saturation is important but, as it does not relate to sample size, it is beyond the scope of the present paper.

FIGURE 1



Note. The dotted vertical line shows the number of beliefs elicited by the 'initial analysis sample'. The "1 2 3" sequence above a line represents the achievement of the stopping criterion: three interviews with no new beliefs emerging.

FIGURE 2



Note. The dotted vertical line shows the number of beliefs elicited by the 'initial analysis sample'. The "1 2 3" sequence above or below a line represents the achievement of the stopping criterion: three interviews with no new beliefs emerging.