Sample Size

In: Sampling and Choosing Cases in Qualitative Research: A Realist Approach

By: Nick Emmel

Pub. Date: 2014
Access Date: September 17, 2018
Publishing Company: SAGE Publications Ltd
City: London
Print ISBN: 9780857025104
Online ISBN: 9781473913882
DOI: http://dx.doi.org/10.4135/9781473913882
Print pages: 137-156

Sample Size

This chapter considers one of the most frequently asked questions, how big (or small) does a sample have to be in qualitative research? I consider the key arguments for sample size in the three sampling strategies of theoretical sampling, purposeful sampling, and theoretical or purposive sampling considered in the first part of the book. I also deal with the practical problem of sample size in qualitative research, which is given little attention in many methodological accounts. The realist sampling strategy considers the ways in which fragments of insight are collected in qualitative research. To ask how big the sample size is or how many interviews are enough is to pose the wrong question. It is far more useful to show the ways in which the working and reworking of relationships between ideas and evidence in the research are a foundation for the claims made from the research.

Large Numbers, Small Samples, Cases

Even studies with apparently large sample sizes are small in qualitative research. The largest qualitative samples do not seem to exceed about 200 units. Savage and colleagues' (2005) sample size of 186 participants and Alan Wolfe's (1996) study with 200 interviewees in the US investigating the experience of being middle class are typical. But even with an apparently large headline number of participants these samples are really rather small. To appreciate how small we need to consider how each sample is divided down. As will be recalled from the discussion in Chapters 5 and 6, Savage and colleagues' sample was sub-divided across four residential areas in Manchester's suburbs; in Wilmslow, 44 participants were recruited (population 30,326), in Cheadle, 43 (population 12,158), in Chorlton, 47 (population 13,512), and in Ramsbottom, 47 household members answered the researchers' questions (population 14,635). Similarly, Wolfe's investigation was in four cities from across the four corners of the United States, in which two areas were selected. Twenty-five individuals were sampled from eight areas; on the East Coast, Boston (Brookline, Massachusetts population 58,732 and Medford, Massachusetts, population 56,173); in the South, Atlanta (Southeast DeKalb County, Georgia, population 691,893 and Cobb County, Georgia, population 701,325); in the Midwest, Tulsa (Broken Arrow, Oklahoma, population 98,850 and Sand Springs, Oklahoma, population 18,906); and on the West Coast in San Diego (Eastlake, California, population 243,916 and Rancho Bernardo, California, population 49,115).

As Wolfe (1998) notes, in a CBS News/New York Times poll conducted in 1992, 75% of respondents considered themselves middle class when asked a question: ‘When presidential candidates talk about the middle class, do they mean people like you?’. That is 194,938,941 people using the US Census Bureau's National Intercensal Estimates (1990–2000). Both

Wolfe's and Savage and colleagues' samples on superficial inspection give the impression of big numbers, but when examined carefully even these relatively large studies are the slightest incursion into the populations investigated.

But these numbers are merely a distraction from the work these samples are doing in the research. Each of these studies builds in strategic comparison to its design. The concern in designing these studies is not how many, but what for. The four areas selected in Manchester were identified because they exemplified ‘core processes and developing typologies around which individuals could be meaningfully linked’ (Savage et al., 2005: 17). The eight areas selected in the US were selected ‘to examine suburbs that on the surface would be as different from each other as possible’ (Wolfe, 1996: 21), but all exemplified certain characteristics that the researchers considered to be typical of middle class suburbs.

Despite each of these studies using variables and categories to decide where and with whom to sample, neither seeks to make claims that are representative of middle class people of Manchester as a whole, nor the apparently huge self-identified middle class of the United States. As Wolfe (1996: 32) observes:

> despite what, for an ethnographic account, might seem like a relatively high number of interviews, two hundred is far too low a number for any kind of survey.

Instead, in each of these cases the concern is less with knowing what proportion of middle class people thought or acted in this way, and what proportion thought and acted in another way. They are more concerned with capturing complexity, nuance, and the dynamics of the lived experience of being middle class and with exposing and exploring critical cases.

Savage and colleagues' (2005) study is almost unique in tabulating how each of the participants contributed towards the rich narrative account they present in their book, how each individual in their sample supported interpretation and explanation. Arbitrarily choosing one among the 186 participants listed in the appendix, D52, a 39- year-old orchestral musician living in Chorlton, contributes his views on why he decided to move to the area, his love of traditional and changing Manchester, his social and work life, and television viewing habits. Wide ranging views, observations, likes, and feelings are mobilised by Savage and colleagues in the service of their analysis and in the production of cases in the research.

The Limits to Explanation and the Number of Cases

Asking an individual about an experience is not to ask them to recount some unique occurrence, but, as Donald T. Campbell (1975) observes, this questioning seeks to provoke a response that allows participants to recount a wealth of experience that relates to the context in which it is described. Participants are chosen because they have this wealth of experience to offer, they have the resource of lived experience to draw on in recounting their story, how these impact on their lives, and the implications of these for social practices. The sample have collected and explained to themselves the collateral experiences of events (March et al., 2003).

As the previous section showed, seeking these accounts from a sample in qualitative research means, inevitably, that from the smallest to the largest qualitative study the sample can only be a fragment. Each of these fragments is a rich elaboration of experiences collected in research. They are not single data points, but detailed stories that elaborate on experience.

In the analytic induction of theoretical or purposive sampling the justification for sample size depends on understanding these experiences in closed social systems. Bertaux and Bertaux-Wiame (1981) (see Table 8.1), for instance, note the relationship between homogeneity of experience and smaller sample sizes and conversely investigations with diverse groups demand a much larger sample. They are conceiving of dissociated critical cases that are brought into conjunction to test the intellectual work in the research. So for instance:

> A single life story stands alone, and it would be hazardous to generalize on the grounds of that one alone, as a second life story could immediately contradict those premature generalizations. But several life stories *taken in the same set of socio-structural relations* support each other and make up, all together, a strong body of evidence (Bertaux and Bertaux-Wiame, 1981: 167 - emphasis in the original).

Bertaux and Bertaux-Wiame are trying to explicate ‘the structural patterns that underlie a given set of social processes’ (1981: 168). Their typical cases are used to test the vitality of theories about structural relations in the bakery trade (see Chapter 3). An adequate sample is defined by informational redundancy, which is the point when the typical cases are filled with information, but this is a redundancy forever constrained in a given moment in a particular set of contingent and dissociated contextual factors.

For realists, these stories, the accounts of experiences and events (see Chapter 4), do not provide the empirical contours for the production of critical cases. They are, instead, opportunities to test and refine ideas, to prove and refute conjectures. Reporting that 1 or 200 cases were collected is not as important as the ways in which insights into events and experiences are used for interpretation, explanation, and claims from research. Realist sampling strategies seek out extensive accounts that expand upon and develop the descriptive baseline of the chosen cases, providing insight into the ways in which phenomena are experienced, explained, perceived, and accepted in particular contexts and circumstances. Acquiring such insights in research means that what are collected are, invariably, large amounts of data. There are practical challenges to collecting these, which I now consider.

The Practical Problem of Data Collection, Analysis, and Reporting

The detail and richness of narrative we seek in qualitative research mean that it is inevitable that qualitative samples are small. As Mason (2002) notes, there is no methodological reason for small sample sizes in qualitative research, but the sample does have to be of a size that can be managed in practical terms. To return to the study conducted by Savage and colleagues (2005) discussed in the last section, they report collecting nearly 1.5 million words of transcript data. To read all of this data is the equivalent of reading one of Charles Dickens' lengthy novels, Our Mutual Friends for instance, nearly four-and-a-half times over. All of this data need to be transcribed, checked, read, coded, and then parts of these transcripts analysed. It is unsurprising that Savage and colleagues (2005) report that initially they were able to do only data mining to address quite focussed research questions. They finished collecting data in 1999. It was only when Mike Savage had a year of sabbatical from 2002–2003 that the bulk of the data could be analysed. In the meantime the research team took up new jobs, took on administrative responsibilities, no doubt spending much of their time considering how researchers in their respective university departments could do more research, and had to care for their children.

Personal and practical limitations on resources must be at the forefront of researchers' minds when considering how many people's accounts, documents, images, artefacts, places, events, research diary entries or whatever we choose to include in the research.

Qualitative samples are invariably small because in collecting rich insight these data will be bulky. Data can be measured in drawers of filing cabinets filled and hard-drive space consumed. Thrift in planning and implementing a sample must always be tempered, however, with an over-arching concern to ensure enough data is collected to gain insight into the complexity of the social processes under investigation. Parsimony is planned in the research from its early stages of conceptualisation. Frugality and richness of account are constantly accounted for in planning, conducting, and reporting cases in research.

All We Have Are Fragments, Experiencing Single Cases Richly

Whether 200 participants or one case is chosen, this choice is made to allow for the interpretation and explaining of social processes. Cases are chosen because they contribute to creatively solving the puzzle under investigation and present as convincing a case as can be mustered with the resources to hand. As noted, Wolfe and Savage are concerned to have enough participants living in each of their predefined areas to say something through comparison about middle class identity, whilst also searching for possible dimensions of difference.

Other researchers may choose quite different tactics. Their concern remains to collect critical cases, and insights that allow for the testing and refining of ideas. It is often reported, for instance, that William Foote Whyte (1993 [1943]) chose one slum and one informant in that slum to investigate the highly organised and integrated social system of Cornerville. His informant, Doc, was an entry point into a diverse, but limited network of the slum.

It is experiences in qualitative research like these that Michael Quinn Patton considers in justifying small sample sizes, even the single case study. For Patton (2002: 245) the key considerations in justifying sample size focus around:

> validity, meaningfulness, and insights generated from qualitative inquiry (which) have more to do with the information richness of the cases selected and the observational/analytical capabilities of the researcher than with sample size.

With this observation realists can agree. Patton's pragmatic approach is a two-sided coin. On the one side judgements are made about how to expend resources - an in-depth enquiry with a small number of sources, even a single case. On the other side of the coin are judgements as to whether these pragmatic choices lead to a sufficiently rigorous and valid account of the subject of investigation. Absent, however, from this account are the drivers to evaluate a case's information richness beyond its empirical insight. Realists will be concerned to understand how each case contributes to the work of interpretation and explanation in the research, and how ideas are tested and refined within cases and between cases.

Phillipe Bourgois (1998) shows how a single case is used to work out the relation between ideas and evidence. He pursued his informant Mikey across the wastelands of East Harlem, New York City on a wet and cold December night to learn about taking heroin at the shooting gallery. The striking feature of this study is its thick description; real experiences, recounted and elaborated open up opportunities to work and re-work ideas.

Bourgois (1998: 64–65) seeks to explain the grim reality of the heroin economy. He talks of the

> multi-billion-dollar drug industry - the only growing equal opportunity employer in America's inner cities since the 1970s … (where) dealers believe with a vengeance in the Great American Dream … the street offers both a real economic alternative and

> also an ideological framework that promises pride and self-esteem.

One case brings together observations as Bourgois trailed around after Mikey, learnt from Doc the manager of the shooting gallery, and watched Slim and Flex shoot up speed balls. These empirical accounts are brought into conjunction with theoretical ideas and an account of the context of the study. Together they produce a trustworthy narrative from the research; the grime, cold, fear, desperation, relief, camaraderie, and hierarchy of the shooting gallery are retold. So too are the mechanisms that position agents in a web of structures, convey norms and inter-relationships, interpret and explain causal powers and liabilities. They place Mikey and his fellow addicts in the wastelands of a New York City suburb on a cold winter's night and in a wider canvas of the political economy of the United States. Observation, interview, overheard conversations, accounts of context, and theories are brought to the study, rejected, appropriated, refined, and revised. These are the case in Bourgois' study; a fragment, but one he chooses and uses richly.

Guided by a similarly realist approach Loïc Wacquant opens up life in the Chicago Projects to the interested reader. His sample is Rickey, or more appropriately Rickey's point of view. He is a convenient sample:

> I met Rickey through his brother, whom I had encountered in the course of my research on the craft of the boxer in Chicago in a gym located at the heart of the ghetto … 'He boxed pro too, he even makin' a come-back, you should interview 'im', suggested Ned. (Wacquant, 1998: 3)

A three hour interview characterised by its 'nervous, up-tempo delivery' (1998: 11) of ghetto living and the life of a hustler gives Wacquant (and most of his readers) an insight into an unknown world, one which Wacquant must position somewhere in time and space:

> Now, it would be a serious mistake to see Rickey as a marginal *curiosa,* an exotic character belonging to a *demi-monde* close to the criminal underworld or liable to an analysis in terms of 'delinquency'. For the hustler, of which he offers a compact personalized incarnation, is on the contrary a *generic figure that occupies a central position* in the social and symbolic space of the black American ghetto (emphasis in the original).

Wacquant's assertion of the generic space Rickey occupies positions the research in an ethnographic body of slum studies. But the purpose of these is not merely to provide empirical coordinates for his sample. The choice of Latin neuter plural and gender unhinged neologism in this account places him a long way from his subject. Wacquant's job as a social scientist is to bring a deeper understanding of the conditions (Bourdieu, 1996) of Rickey's existence. He draws upon and mobilises theory to explain who his case is and how he will be understood in Wacquant's interpretation of hustling and ghetto life.

We have already learnt that Rickey's narrative cannot be explained through the reductionism of ‘delinquency’. Wacquant (1998: 11 - emphasis in the original) places his subject in a wider web of relationships:

> Rickey is not a social anomaly or the representative of a deviant micro-society: rather, he is the *product of the exacerbation of a logic of economic and racial exclusion* that imposes itself ever more stringently on all residents of the ghetto.

In this we have two key features of realist research. First, a rejection of a micro-empirical account in which Rickey is given the tiller-hand. Rickey's interview cannot be read-off as case study, an empiricist representation which attempts, in some way, to convey experience. Even though large parts of the interview are reproduced in the paper, these transcripts do not stand alone. As Bourdieu (1996: 29) argues, ‘[s]ocial agents do not have an innate knowledge of what they are and what they do [their] declarations can, without aiming to mislead, express quite the opposite of what they appear to say’. Realists must steer these weak constructions with strong interpretation and explanation.

The second key feature highlighted by Wacquant is the engagement of the presuppositions and ideas of the social scientist with evidence; a casing strategy to explain what powers and liabilities work for whom, in what circumstances and why. Wacquant's realist interpretation gives Rickey's tales of the ghetto their theoretical moorings far beyond the boundary of the project. This theory explains the causal mechanisms and the scope of the sample.

If the sample in qualitative research is to do all of this work then it cannot be simply described through the definitiveness and precision of a number. The real world is, as Gian-Carlo Rota (1991: 177) observes, ‘filled with absences, with absurdities, with abnormalities, with aberrances, with abominations, with abuses, with *Abgrund* (chasms)’. It is these that are interpreted and explained through our research, which includes the insight of events and experiences from cases, the insiders' perspectives, and the outsiders' understandings.

It is to a brief consideration of this relationship between case and claim that I now turn, before considering the pernicious influence of numbers in descriptions of the sample in qualitative research.

From Cases to Making Claims

Mere empiricism is of little worth, as Boudon (1991) observes. Theories, the claims made from research, are of the middle range. They transcend sheer description or empirical generalisation (Merton, 1968). As seen in the discussion of Wacquant's and Bourgois' realist accounts in the previous section, claims consolidate ideas and evidence into statements that confederate wider networks of theory, yet provide the opportunity to generate hypotheses to be tested through further empirical enquiry. Theories of the middle range are not grand all-encompassing system theories. They are, instead, 'special theories of greater or less scope, coupled with the historically-grounded hope that these will continue to be brought together into families of theories' (Merton, 1968: 48), In other words, theories are fallible, the subject of revision, reinterpretation, and re-presentation. This approach rarely advocates the wholesale ditching of an idea, but is one of constant accretion (Pawson, 2013).

All research, as I argued in Chapter 4, starts with ideas. The problems chosen for research do not come out of the blue but are related to our background knowledge. Investigations are driven by ideas, the sample is chosen using these ideas. From this sample we gain descriptive narratives of events and experiences. This sample may be animate units able to express events and experiences, such as individuals, groups, or organisations. Similarly, our sample may comprise inanimate traces of the relations between structure and agency in documents, photographs, or even the contents of a family's mantelpiece.

Howsoever the sample is composed of sampling units they provide empirical subjective insight. These narratives are, as Margaret Archer (2000: 313) observes, 'about the world and therefore cannot be independent from the way the world is'. For realists there is no direct correspondence between that which we observe, hear, or see, and reality. Social reality cannot be simply captured in description or, for that matter, ideas. It is far richer and deeper than that. Provisional theories about reality are tested, refined, and judged in relation to evidence. This evidence must support conclusions and conclusions must not go beyond what the evidence can support, as Howard Becker (in Baker and Edwards, 2012: 15) argues.

These data may be from samples of one or fewer, as James G. March and colleagues (2003: 469) have it. They 'provide scraps of information about an underlying reality that cumulate, much the way various elements of a portrait cumulate to provide information about its subject'. They are fragments. These are neither independent samples of some universe in a statistical sense, nor will they cumulate to signify a wider population.

In a realist sampling strategy, purposive work allows for a plan to be drawn up of the number of units to be sampled early in the research. These numbers are a plan only, in which ideas, external and internal powers in the research provide an account of the number of observations to be carried out, the interviews conducted, the documents read, and so on. This quota will inevitably change as the research progresses and insight is gained into that which is investigated. As discussed in Chapter 6, the variables used to define quotas presented in early plans are less a blueprint and more a preliminary sketch on the back of an envelope. They are there to be elaborated on as the research progresses.

In common with the inductive strategy, cases in a realist strategy are, as was noted in the previous chapter, subject to processes of repeated and reflexive planning in the research. In answering the question 'how many qualitative interviews is enough?' (Baker and Edwards, 2012: 29) Jennifer Mason's response in this working paper is, 'it depends'. The web of considerations upon which the judgement is made is these:

> a deep exploration of how processes work in particular contexts, under certain sets of circumstances, and in particular sets of social relations … a more interpretative and investigative logic … so that you build a convincing analytical narrative based on the argument that you have explored the process in its richness, complexity and detail, and that you have understood the contingency of different contexts.

What Mason argues for here is strongly interpretative. The quota of contexts, circumstances, and social relations will start with a number for practical reasons. Cases will be described and re-described throughout the research. Ideas are brought into play with evidence through its collection and interpretation. Cases bear these characteristics as the relation between ideas and evidence are worked out through the research.

Yet frequently the breaks are put on this reflexive process and the sample is quantified. It is asserted that a particular size of sample is adequate to investigate a research question. Often this insistence on a larger sample size is imposed by external liabilities and powers: ethics review boards, journal reviewers (an example of which I discuss shortly), grant proposal reviewers, examiners, and academic supervisors who are worried by all of the above. A pseudo-quantitative logic is imposed that assumes a large number sample is more reliable towards producing trustworthy findings from research.

Pronouncements from external powers to increase sample size can be explained through the imposition of a dominant ideology of quantitative reasoning. Yet, it is surprising, even intriguing, how many qualitative methodologists advocate an acceptable sample size for qualitative studies. It is to this allure of numbers that I now turn.

The Allure of the Number *N*

There are no guidelines, tests of adequacy, or power calculations available to establish sample size in qualitative research. Yet qualitative researchers persist in using a mathematical notation *(n)* to describe their sample size. This is emphasised in Table 8.1 which reports how qualitative researchers have found it necessary to state that particular kinds of studies across qualitative idioms require a particular sample size or range. In most of these examples no evidence is provided for the chosen range. The impression gained from reading the accounts in which these assertions are made is that this sample size worked in a specific study undertaken to investigate a particular phenomenon, with a particular population, in a particular setting (see for instance Morse (1994) and Creswell (1998) in Table 8.1). Given that replicating studies to account for these dimensions is highly unlikely, generalising from these numbers and applying them is not productive to the research of others. The guidance offered in Table 8.1 has little if any value in determining sample size in qualitative studies.

Table 8.1 The limited value of stating *n* in qualitative research studies

Author	**Sample size (*n*)**	**Notes**
Bertaux and Bertaux-Wiame (1981)	15–30	Depends on the variety of structural experience - based on research with bakers (homogeneous group) bakery owners (heterogeneous group)
Kuzel (1992)	6–8	Homogeneous sample (assertion, no evidence)
	12–20	Heterogeneous sample - ‘when looking for disconfirming evidence or trying to achieve maximum variation’ (assertion, no evidence)
Morse (1994)	6	Phenomenological studies (assertion, no evidence)
	35	Ethnographies, grounded theory studies, ethnoscience (assertion, no evidence)
	100–200	Qualitative ethology (detailed study of behaviour) (assertion, no evidence)
Creswell (1998)	5–25	Phenomenological studies (assertion, no evidence)
	20–30	Grounded theory studies (assertion, no evidence)
Bernard (2000)	36	Most ethnographic studies seem to be based on this number (assertion, no evidence)
Guest et al. (2006: 79)	12	‘For most research enterprises … in which the aim is to understand common perceptions and experiences among a group of relatively homogeneous individuals, twelve interviews should suffice.’
Adler and Adler in Baker and Edwards (2012)	30	A good round number to aim for (a practical consideration and acceptable to external powers)

12	A student's one semester study (a practical opportunity to practice qualitative research skills)
20	A student's two semester study (a practical opportunity to practice qualitative research skills)

The Limits of Theoretical Saturation

The most commonly mentioned justification for a stated sample size in qualitative research, Mark Mason (2010) assures us, is theoretical saturation. This is based on his extensive review of PhD studies using qualitative methods. Theoretical saturation is described across the approaches to grounded theory discussed in Chapter 1. Corbin and Strauss (2008) suggest that less than 5–6 interviews are not enough to achieve saturation but do not identify quite how large a sample should be. Mason (2010) in his investigation of 560 PhD studies found that a mean average sample size of 31, but with a widespread distribution (standard deviation 18.7). There was, Mark Mason went on to note, a preponderance of studies including 10, 20, 30, and 40 participants.

It is intriguing to ask why the mean average is 31. One reason may be that 30 degrees of freedom is when Student's t-distribution used for small *n* samples in statistics approximates to a normal distribution, an unwritten pseudo-quantitative logic that 30 is a small sample in statistical research so it will do for qualitative research with its small sample sizes. Of course this is just lateral thinking, but of more concern in a discussion about sample size is Mason's observation that PhD students are not adhering to guidelines for theoretical saturation. He contends that the problem with these guidelines is their elasticity. Yet, Greg Guest, Arwan Bunce, and Laura Johnson (2006) felt able to devise a sophisticated experiment to quantify when theoretical saturation is achieved suggesting that these guidelines can be rigorously applied.

The definition of theoretical saturation Guest and colleagues (2006: 65) use is ‘the point in data collection and analysis when new information produces little or no change to the codebook’. Their experiment is designed to gain a ‘reliable sense of thematic exhaustion and variability within … the data set’.

This experiment asks several key questions of theoretical saturation. How many interviews are needed before no new codes are discovered? How many interviews are needed before codes are filled and no further empirical data is needed? And finally, what value is there to the study through using comparative groups? The experiment addresses the key dimensions of theoretical saturation as discussed by Glaser and Strauss (1967) and which Glaser (2001: 191) summarises:

> Saturation is not seeing the same pattern over and over again. It is the conceptualisation of comparisons of these incidents which yield different properties of the pattern, until no new properties of the pattern emerge. This yields the conceptual density that when integrated into hypotheses make up the body of the generated grounded theory with theoretical completeness.

Guest and colleagues conduct their experiment within a study to investigate the accuracy with which women who fall into a high risk group of acquiring HIV infection talk about their sexual encounters. Thirty sex workers were recruited in Ibaban, Nigeria and Accra, Ghana, from three high-risk sites, a red light area, a hotel, and a hostel. The criteria for selection were that the women were eighteen years of age or older; had vaginal sex with more than one male partner in the previous three months; and had vaginal sex three or more times in an average week. Each woman was asked identical questions in the same order from an interview guide, but the interviewers were encouraged to ask a number of sub-questions if particular issues were raised. Interviewers were also encouraged to probe key themes.

Working with batches of six transcripts (the smallest recommended sample size these authors identified in the literature), Guest and colleagues audited the newly created codes and changes to existing code definitions. They measured the frequency with which codes were applied. Starting with the interviews collected in Ghana, these sedulous researchers coded the interviews six at a time until all 30 were coded, then moved on to add the analysis of the codes collected from Nigeria until all 60 interviews had been audited.

The final codebook was made up of 109 content-driven codes, of which 80 (73%) were identified in the first six interviews. A further 20 (18%) were identified in the next six. After analysis of the first 12 interviews, 92% of the codes had been discovered. After reviewing all thirty interviews from Ghana the researchers completed their codebook. They moved on to the Nigerian data, adding one more substantive code and developing four of the codes as variations on existing codes. 'Two of the four new codes were needed for the unique sub-group of campus-based sex workers' (Guest et al., 2006: 66). This group neither referred to themselves as sex workers, nor to their sexual partners as clients. The codes were modified to account for the different ways in which this sub-group talked about themselves and clients.

The first part of this audit revealed the frequency of coding. The second investigated how codes changed as the research progressed. In all, 36 changes were made to codes during the research; eleven percent in the first round of analysis, the largest number of change in the second round (17 changes, 47%), and 20 percent of changes in the third round. By the time 18 interviews had been audited, 78% of the changes had been made to codes.

A third test sought to identify the thematic importance of each code, through identifying the internal consistency of participants' accounts. After the first 18 interviews, participants were consistently addressing all the most important themes identified in the research. The most frequently discussed themes were elaborated on in the early interviews. Thirty-six codes were mentioned frequently by participants, of these 34 (94%) were mentioned in the first six interviews, and 35 (97%) after 12 interviews. Guest and colleagues conclude that very little was missed in the early stages of analysis.

This experiment investigates two of the three themes Glaser and Strauss (1967) consider with reference to theoretical saturation, the empirical limits of the data in producing codes, and the depth of each category. Their experiment meets Glaser's demand that theoretical saturation should include approaches to coding and refining of these codes to aid conceptual density. But Guest and colleagues (2006) did not seek to develop different slices of data through using different research instruments, the third element of Glaser and Strauss's schema for theoretical saturation. The semi-structured interview instrument applied by Guest and colleagues was, in fact, quite rigid and was applied with all 60 of their participants. This has important implications for any interpretation of the results from this experiment and its application to other studies.

The important finding from this research is that after 12 interviews 100 of the 107 codes (93.45%) are discovered and 97% of the changes are made to these codes. In an objectivist grounded theory logic of theoretical saturation a very large proportion of empirical findings have emerged, and, therefore, theory discovered. Guest and colleagues (2006) question whether there is much value in doing more than 12 interviews. The final 38 interviews produce only marginal returns of theory discovery for a large expenditure of resources.

At first sight, mobilising a comparison does not seem particularly fruitful towards identifying new codes or saturating existing codes either. This experiment does not report on differences between the three high-risk sites used to access participants in each country. The emergence of new codes and the refining of codes were very small when the Nigerian data were added to and compared with the Ghanaian data. Part of the explanation may rest in the similarity of experience between sex workers in the two countries, although Guest and colleagues (2006) do highlight apparently important differences between the two country groups. This lack of discovery is intriguing and does raise questions as to why further insights that produced new codes and refined existing codes did not happen. Part of the answer may lie in the openness and theoretical sensitivity of the researchers in each country. As Guest and colleagues (2005) observe in another paper, there was an important methodological difference between the interviews conducted in Ghana and Nigeria. Despite the interviewers in both countries using the same semistructured interview instrument and receiving the same training, the ‘Nigerian interviewers did not probe responses as readily, rendering significantly shorter responses’ (2005: 287).

N = 12, How Reassuring?

Guest and colleagues' paper, *How many interviews are enough? An experiment with data saturation and variability* (2006) has become reassuringly useful to some qualitative researchers seeking a justification for sample size. I have heard at least one PhD supervisor observe that she tells her students their sample size should be *n*=12, based on this paper. A recent blog in Methodspace makes a similar argument, referring to one of the external powers discussed above:

> I and a recent doc. student grad. have had difficulty with some journal review boards with the small sample size of qualitative research. I have tried to communicate the point that saturation is more important than the sample size. Does anyone have any references that may discuss sample size in qualitative research? …
>
> My student had 4 participants, 2 rounds of interviews and member checks with each for a total of 12 interviews. Guest, Bunce, and Johnson (2006) found with their study that involved 60 interviews theme saturation was achieved after 12 interviews. I use[d] their study as [it] supported my student's decision.
>
> <http://www.methodspace.com/forum/topics/sample-size-and-number-of?id=2289984%3ATopic%3A12428&page=1#comments>

Guest and colleagues (2006) emphasise considerable care should be exercised when relying on the authority of their experiment to justify theoretical saturation and sample size. First they note that such small samples might well work if those with whom research is being done share common experiences. Many of the women in their study experienced fear of being exposed as sex workers by the media, and according to the authors they work in very similar contexts. The samples' experiences are relatively homogeneous. The study design reinforces this homogeneity through its fairly limited research questions. It enquired into a narrow range of experiences, with, as discussed above, each participant asked a similar set of questions. Guest and colleagues (2006) consider that without these factors - a relatively homogeneous substantive experience that is widespread in the target population, and a narrowly-focussed and prescribed method - the ability to achieve theoretical saturation in a grounded theory methodology cannot be reached. Their finding, that *n*= 12 interviews is sufficient, is ‘not applicable to unstructured and highly exploratory interview techniques', where saturation 'would be a moving target, as new responses are given to newly introduced questions.' (2006: 75).

As soon as the phenomenon under investigation is recognised to be dynamic, contingent, and best explored through detailed and in-depth investigation, then the principles of theoretical saturation demonstrate considerable weaknesses.

There is also compelling evidence of a feedback mechanism at play between a narrowness of approach and a narrowness of theoretical discovery through coding in this experiment. The lack of new codes emerging from the very different sex workers in Nigeria is surprising. Part of this might arise from the lack of training among the Nigerian researchers, as noted. But a further element could well be what Dey (2007) describes as theoretical sufficiency, in which categories suggested by data rely on researchers' conjecture. They stop short of coding all data through applying a strictly enforced search for themes and codes, which in turn foreclose possibilities for innovative and creative interpretation and explanation from the events and experiences being coded. The emphasis is on coding as simplifying or reducing complexity rather than complicating data through its conceptualisation, 'raising questions, and providing provisional answers about the relations among and within the data', as Amanda Coffey and Paul Atkinson (1996: 31) suggest should happen in any process of data analysis.

Guest and colleagues (2006) take the simplifying approach to coding. The numbers game, how many interviews, observations, focus groups, documents, units, or whatever are enough, rests on an empiricist and positivist assumption that the insight from qualitative research is the sum of the parts, appropriately collected, reduced, and processed to provide descriptive answers.

Charmaz (2006), who, as argued in Chapter 1, seeks to distance her account of a constructivist grounded theory from its positivist roots, contends that objectivist methodologies of grounded theory have the potential to 'force data into preconceived frameworks … (it) takes the focusing inherent in grounded theory and renders it directive and prescriptive' (2006: 115). As she also observes, Guest and colleagues' result of 12 interviews 'may generate themes, but not respect' (Charmaz in Baker and Edwards, 2012: 21).

Respect, she argues, comes from recognising the wealth of insight that should come from mixed qualitative research methods. Beyond these observations Charmaz does not distance herself from the micro-empiricism of grounded theory, however. Her solution to the foreclosure of interpretation inherent in objectivist grounded theory is to return to the data, to recode it towards defining new leads. Researchers should, in Henwood and Pidgeon's terms:

> avoid being wedded to particular theoretical positions and key studies in the literature

> in ways that overtly direct ways of looking and stymies the interactive process of engagement with the empirical world studied. Theoretical agnosticism is a better watchword than theoretical ignorance … (2003: emphasis in the original).

Theory falls out of the data in this account of its co-construction. It is a re-framing of Glaser and Strauss's insistence on openness and theoretical sensitivity, underwritten by an assumption that it is the number of engagements with the empirical world that are key, even if the emphasis is on depth of engagement as well as breadth. Saturation, whether configured through an objectivist or constructivist approach to grounded theory insists that size of the sample matters. Both approaches assume that theoretical saturation arises from the neutrality of the researcher in their coding of data, and from this coding we can determine a number, a count of data points where as Morse (1995: 148) observes in the objectivist tradition, there is 'enough data to build a comprehensive and convincing theory', and Charmaz (2006: 114) contends, from a constructivist standpoint that 'a study of 25 interviews may suffice for certain small projects but invites scepticism when the author's claims are about, say, human nature or contradict established research'.

Making the Sample Work

In 1659, Blaise Pascal observed that 'nature confutes the sceptics, reason confutes the dogmatists' (translated and quoted in Lakatos, 1976: 54). Put more plainly, realists will always find their explanation of nature wanting, empiricists will find their methods inadequate for the task.

Drawing on more recent debates, the empiricism of grounded theory and theoretical sampling was the focus of Chapter 1. Grounded theory hinges on concerns that the procedures to transform empirical data points to theory through coding are correct. Theoretical saturation, similarly, appears to offer a method through which claims can be made for the adequacy of a sample size in a study. As I have shown, however, such claims to saturation can only be made through simplifying and reducing the complexity of insight.

The selection of information rich cases is the preferred strategy of purposeful or judgemental sampling strategies. These empirical cases can be evaluated for their rigour. They are sufficient for the enquiry of which they are part. The number of cases is part of this pragmatic consideration too. Sample size is evaluated in a similar way to the choice of which of the 14+1 strategies of purposeful sampling should be applied in a particular study (see Chapter 2), which strategy most adequately reflects observed reality in the most convincing way for its intended audience.

The strong interpretation of inductive strategies of theoretical or purposive sampling takes the discussion about sample size in yet another direction. The critical case combines intellectual work with the empirical contours in the research. As was discussed in Chapter 3, the task is to build explanation of how social processes work in particular contexts, within certain social relations, and under particular circumstances. This deep exploration interprets the conditions under which causal relations operate, through strategic comparison of cases, whether considered typical or negative. The case is achieved through its closure, dissociating it from anything other than the causal powers that are being investigated. In this way sample size is always justified in relation to the explanation from the research. There is sufficient information to interpret the 'experience' as a 'complex interactional process involving many happenings and events' (Lindesmith, 1968: 13 - emphasis in the original), any more insight would be redundant. Induction is found wanting in its explanation of nature. Nature is an open system. For realists, samples are always fragments drawn from this open and stratified system.

The focus of any realist justification of the cases chosen in the research is on the adequacy (O'Reilly and Parker, 2012) of the data collected. These justifications are constantly informed by preconceived theory that shapes the choices we make about whom or what to sample. The task is to demonstrate how the fragments available are used towards explanation and interpretation. The consideration in a realist sampling strategy is what work we can set the sample to do in the research to test, refine, and adjudicate between ideas.

There is no reasonable methodological way in which realist qualitative researchers can tell you the eventual number of cases in advance. If they provide a number it is likely that they are complying with external liabilities and powers from institutions, ethics review boards, funding bodies, and/or editors of journals. They are obliged to assume that those who review their research are beguiled by a pseudo-quantitative logic. They hold that a largish number (generally ≥ 30) of interviews, focus groups, or whatever instruments chosen, equates with a trustworthy outcome from the study.

Quite often a number is stated in research plans and proposals. This can be interpreted as a sample size. It is not. This number is an estimate, based on evaluation of human and financial resources available. It will include a calculation of the number of times research instruments are used and an estimate of the time to recruit participants, apply the instrument, and deal with the data generated, use these in interpretative and explanatory work in the research, and disseminate the knowledge generated. In addition, any count of potential units to be sampled will include purposive work, the incomplete accounts of who or what would be useful to do research with, a given number of units that have a given set of characteristics chosen towards testing and refining theories of the middle range.

While this work is often done, the sample size stated at the beginning of the study will not be reflected in the final account of the cases. This casing is done for two reasons. First, the descriptive baseline of units chosen to be studied is continuously extended throughout the research. That which we thought we sampled initially is not what we realise we cased eventually. Secondly, we will work and rework the relationships between ideas and fragmentary evidence throughout the research. This is the methodological work of casing. A number can be assigned, eventually, to the number of cases in any piece of research, but it is not the number of cases that matters, it is what you do with them that counts. Sample size are frequently used to determine the quality of both qualitative and quantitative research design, as Emma Uprrichard (2013b: 7) observes. But in realist research this criteria is rendered meaningless 'without further explanation as to what, how and why [it] may matter in the first place.'

In the introduction to this book I suggested choosing cases in qualitative research is better understood through inverting the traditional account of sampling. In realist qualitative research the sample can only be weakly elaborated beforehand. It is at best a weak construction which raises consciousness about the cases in the research and why they are chosen. Little of value can be said about what these will represent at the outset, but a great deal will be learnt as the research progresses. Descriptions will change from fragile accounts using variables and categories to strong interpretations and explanations incorporating causal mechanisms, contexts and outcomes.

There is no methodological reason for ensuring that every person or thing from a predefined population has the same chance of inclusion in an investigation. A strategy for choosing cases that is random or stratified may be designed into the study for particular theoretical reasons. But, like strategies, cases are chosen because they can be worked and reworked through the research. Choosing cases provides opportunities to interpret and explain social phenomena, and the powers of these things that exist independent of our accounts of them, as best we can.

http://dx.doi.org/10.4135/9781473913882.n9