Qualitative Research in Sport, Exercise and Health

To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales

Virginia Braun and Victoria Clarke

Abstract

The concept of data saturation, defined as 'information redundancy' or the point at which no new themes or codes 'emerge' from data, is widely referenced in thematic analysis (TA) research in sport and exercise, and beyond. Several researchers have sought to 'operationalise' data saturation and provide concrete guidance on how many interviews, or focus groups, are enough to achieve some degree of data saturation in TA research. Our disagreement with such attempts to 'capture' data saturation for TA led us to this commentary. Here, we contribute to critical discussions of the saturation concept in qualitative research by interrogating the assumptions around the practice and procedures of TA that inform these data saturation 'experiments' and the conceptualisation of saturation as information redundancy. We argue that although the concepts of data-, thematic- or code-saturation, and even meaning-saturation, are coherent with the neo-positivist, discovery oriented, meaning excavation project of coding reliability types of TA, they are not consistent with the values and assumptions of reflexive TA. We encourage sport and exercise and other researchers using reflexive TA to dwell with uncertainty and recognise that meaning is *generated* through interpretation of, not excavated from, data, and therefore judgements about 'how many' data items, and when to stop data collection, are inescapably situated and subjective, and cannot be determined (wholly) in advance of analysis.

Key words: Codebook; coding reliability; data adequacy; information power; information redundancy; interpretation; meaning; reflexive; sample; theoretical saturation

"Of course we saturate, but…"

This quotation was the start of a question about determining sample size that a postgraduate student asked one of us in an 'ask me anything' session on qualitative health

research. The phrasing of the question – in the classic disclaimer format (e.g. van Dijk, 1992) – is revealing. It signals saturation as both taken-for-granted, unquestioned, and maybe even unquestionable, as a criterion for determining sample size in qualitative research ("of course we saturate"), but as also failing to provide all the answers ("but…"). The confused student never got to finish her question; Victoria interrupted to challenge the taken-for-granted status of saturation, something we interrogate here in this paper. We aim to contribute to critical discussions of the saturation concept in qualitative research, and particularly the notions of code-, data- or thematic- saturation often employed in thematic analysis (TA) research, including research citing the reflexive TA approach we have outlined (Braun & Clarke, 2006, 2019). We home in on a cluster of papers that have sought to provide *concrete* guidance for determining how many interviews or focus groups are enough to achieve some degree of 'information redundancy' or data saturation in TA research, in *advance* of data collection and analysis, by effectively 'operationalising' the saturation concept. We question the assumptions underlying the procedures and practices of TA, and the conceptualisation of data saturation, in these papers. This paper extends our critique of practices around determining sample size in TA; elsewhere we have questioned the coherence of statistical models for determining sample size in TA research in advance of data collection (Braun & Clarke, 2016). We continue to use the language of 'sample size' in this paper, despite feeling that this, itself, risks evoking the very neopositivist-empiricist framings we are calling into question.

Saturation as information redundancy

The concept of saturation, often broadly and loosely defined as information redundancy (Lincoln & Guba, 1985), the point at which no new information, codes or themes are yielded from data, evolved from the more tightly conceived notion of theoretical saturation in grounded theory. Theoretical saturation has been defined as the point at which the properties of categories and the relationships between categories are comprehensively explained so that a theory can arise (Morse, 2015). Theoretical saturation is inextricably linked to the practice of theoretical sampling and concurrent practices of data collection and analysis in grounded theory (Hennink et al., 2017; Morse, 2015; O'Reilly & Parker, 2012; Saunders et al., 2017; Vasileiou et al., 2018), meaning that theoretical saturation *cannot* be determined in advance of data collection and (at least some) data analysis. Dey (1999: 257)

described saturation as an "unfortunate metaphor"; it suggests completeness of understanding and a determinable, fixed point for stopping data collection. Some have argued that this was never the intention of the original grounded theory proponents, Glaser and Strauss (1967; see Nelson, 2016; Saunders et al., 2017), and that the language of 'no new' does not capture the nuances of their conceptualisation of theoretical saturation and the refinements of the concept in both their and others' later work (Low, 2019). However, it is clear grounded theorists' statements around repetition and redundancy – "no additional data" (Glaser & Strauss, 1967: 61) and "no new properties" (Charmaz, 2006: 189) – *have* informed the widespread conceptualisation of saturation as information redundancy (Low, 2019).

Dey suggested the phrase *theoretical sufficiency* as an alternative to saturation, to capture the notion that data collection stops when the researcher has reached a sufficient or adequate depth of understanding to build a theory. Nelson (2016) similarly suggested *conceptual density* or *conceptual depth*. From this perspective, theoretical saturation is as much, or even more, about the *quality* of data collected – their richness, depth, diversity and complexity, what can be glossed as data or sampling adequacy – as it is about simply the *quantity* of data collected (Fusch & Ness, 2015). However, in much wider qualitative discussion, saturation – explicitly or implicitly conceptualised as 'no new information' – appears often as a shorthand simply to rationalise and validate the sample size. We use the term *data saturation* in this paper to capture such widely-used information redundancy conceptualisations of saturation (e.g. reflected in notions of code and thematic saturation).

Data saturation – a qualitative research requirement?

The concept of data saturation (especially as or for validity) is firmly embedded within (certain) qualitative research logics. For Constantinou et al. (2017), data saturation is "the flagship of validity for qualitative research" (p. 585), a criterion that "meets with the ontological and epistemological foundations of qualitative research" (p. 583). The opening line of a paper on sampling and qualitative research for PhDs states that "a number of issues can affect sample size in qualitative research; however, the guiding principle should be the concept of saturation" (Mason, 2010: para 1). (Data) saturation has also been identified as the most commonly evoked justification for sample size in qualitative research in the health domain (Vasileiou et al., 2017). Many widely acknowledged 'titans' of qualitative health and

applied research (e.g. Chamberlain, 2010; Morrow, 2005; Morse, 1995, 2015; Sandelowski, 1995) are frequently cited as proponents of saturation, and as evidence for the relevance of the concept for (all) qualitative research. *We* are even cited as recommending that a minimum of 12 interviews are required "to reach data saturation" (Picariello et al., 2017: 386) – though we do not say anything like this in the source cited (Braun & Clarke, 2013).

(Data) saturation as criteria for quality also features in 'quality checklists,' such as the 32-item *Consolidated Criteria for Reporting Qualitative Research* (COREQ) checklist for interview and focus group research (Tong et al., 2007), compiled from 22 checklists, and widely used in health research. Item 22 is "data saturation … Was data saturation discussed?" Similarly, the *Critical Appraisal Skills Programme* 10-item checklist for qualitative research (CASP, 2018) suggests readers consider if the researcher has discussed saturation of data. The American Psychological Association Publications and Communications Board Working Group's *Journal Article Reporting Standards for Qualitative Research* (JARS-Qual) recommend authors discuss the rationale for stopping data collection and offer saturation as an exemplar (Levitt et al., 2018). In this way, saturation – often not defined – becomes normalised as conceptual tool and implicit evidence of good practice, for qualitative researching. Leading to a situation where, for the student quoted above, a disclaimer format is deployed when asking a question suggesting saturation might not be the full answer.

'Evidencing' data saturation for TA research

Data saturation has also become closely linked to TA. Greg Guest, a proponent of one type of approach to TA, has described data saturation as the "gold standard" for determining sample size in purposive samples in qualitative health research (Guest et al., 2006: 60; see Guest et al., 2012). Setting aside for now a failure to explain *why* data saturation is the gold standard – something we are troubled by – Guest et al. (2006) and Constantinou et al. (2017) are among a number of authors who have sought to (to some extent) 'operationalise' the concept of data saturation in TA (and its close cousin qualitative content analysis), to provide practical guidance on estimating sample size in advance of data collection (see also Ando et al., 2014; Coenen et al., 2012; Francis et al., 2010; Guest et al., 2016; Hagaman & Wutich, 2017; Hancock et al., 2016; Hennink et al., 2017, 2019; Namey et al., 2016). In the wider methodological context, concrete sample guidance around 'how many is enough' –

based on ‘data saturation’ – is seductive, especially when the number is relatively small and therefore eminently achievable, particularly when time and resources are tight.

Guest et al. (2006) defined saturation as: 1) *data* saturation – “the point in data collection and analysis when new information produces little or no change to the codebook” (p. 65), with changes consisting of the addition of new codes and refinements of code definitions; and 2) as “thematic exhaustion” (p. 65) – the point at which no new themes “emerge” from data. This definition is consistent with the conceptualisation of saturation as information redundancy. Using data from an interview study, Guest et al. found that 94% of what they call high frequency codes, codes applied to many interview transcripts, were identified within the first six interviews and 97% after twelve interviews (they reviewed theme development and their codebook after every sixth interview, hence the multiples of six; no rationale was given for this). Thus, “data saturation had for the most part occurred by the time we had analysed twelve interviews” (p. 74). Guest et al. contextualised this claim, in relation to the fairly narrow objectives of their study, the relatively homogenous population and the degree of structure to the interviews (similar questions were asked of all participants), and queried the ‘generalisability’ of their findings.

Unfortunately, their nuancing is often lost when their paper is referenced as evidence that it is possible to achieve (data) saturation in twelve or even six interviews (or other data items), in TA research, including research citing our approach – an approach quite different from Guest et al.’s (2006). As an example, in research assessing the thematic content of videos, Marshall et al. (2018) deployed (data) saturation as the justification for the size of the sample selected for TA. They defined saturation as “the point at which no new themes are emerging from the data” (p. 608), and, citing Guest et al., noted that “data saturation was assessed upon viewing the eighth video and again upon viewing the twelfth video, as this is regarded the minimum sample size for reaching data saturation” (p. 608). In another example, Schweitzer et al. (2018) seemed to use saturation – they used the term “theoretical saturation” to refer to no new information – to determine sample size *during* data collection: “Recruitment continued until theoretical saturation had been achieved at 12 participants; this is consistent with Guest, Bunce, and Johnson (2006) who found that data saturation in thematic analysis occurred at approximately 12 interviews” (p. 110). And, from the field of exercise research, with saturation defined around “no new emergent themes” in

transcripts, Eynon et al. (2018: 1479) reported: “through using a set of nine interviews, data saturation occurred after eight analysed transcripts, with the final transcript used to further substantiate the themes outlined (Guest et al., 2006).”

Some researchers report engaging in simultaneous data collection and analysis, connected to data saturation:

> Data analysis was intertwined with the interview process from the beginning. This analysis helped the interview process, provided new topics and enabled detection of data saturation. Data saturation, meaning that no new codes emerged from the analysis, was reached after 24 interviews. Two additional interviews were performed in which data saturation was confirmed (Bragaru et al., 2013).

Data saturation, here defined as no new codes, was determined *during* data collection and *from* data analysis. Other researchers seem to determine data saturation on the basis of their *impressions* of the data during or after data collection. For example, “the principal investigator reviewed the audiotaped and transcribed notes throughout the study to monitor saturation, ending data collection when saturation was reached in both subsamples. Interviewers also discussed saturation and key findings together after each interview session” (Underhill et al., 2015: 670).

These examples illustrate the ways data saturation – variously defined as no new information, codes or themes (mentions of no further code and theme *refinements* are far less common) – has been used to determine sample size at various points in the TA process: during data collection/prior to analysis, following what might be called data familiarisation, and during data analysis itself (which may or may not be independent of data collection). Within such claims, (data) saturation is commonly referenced a way that leaves unclear how exactly it was defined and indeed determined (Bowen, 2008; Malterud et al., 2015), as if it is self-explanatory (as in the widely used ‘the data were saturated,’ or “a point of saturation was achieved” [Marshall et al., 2012: 19]). This suggests to us that the concept of data saturation is used, at least partly, and perhaps wholly in some instances, as a *rhetorical* device, rather than a considered methodological *practice*, an orientation to and deployment of a concept often perceived to act as a concrete and definitive guarantor of the appropriateness of sample size (Morse, 2015).

Other data saturation 'experiments' for or with TA have concluded that data saturation can be achieved in similarly small(er) samples (of interviews/focus groups). For example, Constantinou et al. (2017: 582) claimed that "all possible themes" were found after interview 7. Francis et al. (2010) aligned with Guest et al. (2006) in claiming that 10 + 3 interviews was "a fairly effective guide" (p. 1241) for sample size in theory-based analysis, comparing this to the 0.05 significance criterion in quantitative research. The +3 referred to the number of interviews without any additional material, they claimed as needed to confirm the stopping criteria. Ando et al. (2014) reported that 12 interviews provided all of the themes identified and most of the codes from a sample of 39 interviews. Thus, they concluded that 12 interviews "should be a sufficient sample size for thematic analysis where higher level concepts are concerned" (p. 7). They illustrated their understanding of higher-level concepts with an example – the effect of general physiological symptoms on well-being – and contrasted this with an example of a lower level concept (a list of sensory symptoms and their distinct differences). Hagaman and Wutich (2017), drawing on interviews collected from four sites and a total sample of 132 respondents, focused on "thematic saturation" and how many interviews it took to identify (site-specific) "common themes" and (cross-cultural) "metathemes" *three times* – three because "this is the minimum number needed to fully understand and define the themes" (p. 27). They identified that 16 interviews or fewer were enough to identify common themes from relatively homogenous, site-specific, groups (but 20-40 interviews were needed to reach saturation for metathemes).

So, with the exception of "metathemes" (Hagaman & Wutich, 2017: 26), recommended sample sizes to achieve data saturation within TA have ranged from 6-16 interviews, depending on the specific characteristics of the research and the degree of data saturation required. And, indeed, with where and how data saturation is evidenced. But the concrete guidance provided by these papers often seems to rely on rather arbitrary and largely unexplained criteria, for what counts *as* data saturation – saturation is, ironically, rather poorly 'operationalised' in these 'experiments'. Is a theme 'saturated' after three instances have been identified (Hagaman & Wutich, 2017)? Is a code 'saturated' when *one* instance has been identified? That assumption *seems* evident in all of the papers, with the exception of Hennink et al.'s (2017, 2019) concept of meaning saturation; they suggested the necessity of distinguishing between code- and meaning-saturation, and different *types* of codes, and

offered a refinement of Guest et al.'s (2006) saturation 'experiment'. Drawing on data from a 25-interview study, Hennink et al. (2017) critiqued Guest et al. (2006) for prioritising *prevalence* of codes and themes, rather than *meaning*, and the development of a full understanding of phenomena. Indeed, Hennink et al.'s (2017) conceptualisation of saturation returns us closer to the original grounded theory conceptualisation as *theoretical* saturation, focused on the facets of a concept (or a theme). Hennink et al. (2017) defined *code* saturation as "the point when no additional issues are identified and the codebook begins to stabilise" (p. 594), which encompassed both the refinement of existing codes and the addition of new codes. They distinguished between four *types* of codes: 1) inductive (content driven and raised by participants); 2) deductive (researcher-driven and developed from the interview guide); 3) concrete (capturing explicit, definitive issues); and 4) conceptual (capturing abstract constructs). *Meaning* saturation was defined as "the point when we fully understood issues, and when no further dimensions, nuances, or insights of issues can be found" (p. 594). Similar to Guest et al. (2006), they reported that code saturation was reached after nine interviews: the first interview contributed 53% of codes and 75% of high prevalence codes, "thus, by nine interviews, the range of common thematic issues was identified, and the codebook had stabilized" (p. 598). High prevalence concrete codes were identified and reached meaning saturation earlier, in nine interviews or fewer. However, low prevalence conceptual codes were identified later, and required between 16-24 interviews to reach meaning saturation, or did not reach meaning saturation. Despite their more nuanced take, Hennink et al.'s (2017) study still suggested that various degrees of (meaning) saturation are possible in a sample of 25 interviews, which incidentally is around the mean sample size for interview studies identified in several reviews (e.g. N=21/23 in Clarke & Braun, 2019; N= 31 in Mason, 2010).[1]

The criteria for (data) saturation across these 'experiments' appear to rely on an understanding of codes and themes as entities that pre-exist analysis (to some extent), that reside *in* data, that codes and themes are fixed and unchanging, and that instances of a theme are interchangeable, rather than being the product or output of analysis and representing situated and contextual interpretations of data (Sim et al., 2018a) – which is how we conceptualise themes in reflexive TA (Braun & Clarke, 2019). Even Hennink et al. (2017, 2019), who distinguished between code and meaning saturation, seemed to regard

meaning as 'in' data, awaiting identification. This conceptualisation also suggests to us that in these (data) saturation 'experiments,' codes capture relatively slight observations, or insights about the obvious or concrete – things that are somewhat 'easily' evidenced. But, as we will argue later, it can (and maybe *should*) be more complex than that.

Regardless of the particular definition of saturation used, these studies collectively demonstrate an implicit and explicit lauding of (data) saturation as a gold standard for determining interview sample size in TA research, and something to be aspired to. And, with the conclusions they have reached, it is something apparently achievable in the sample sizes typical of (much) published and doctoral research. But there is far more at play and at stake in considering saturation in (and beyond) TA.

Questioning saturation

There is, concurrently, increasing critical discussion related both to the imprecise use of saturation (e.g. Bowen, 2008; Fusch & Ness, 2015; Kerr et al., 2010; Mason, 2010; Saunders et al., 2017; Vasileiou et al., 2018), and to its often-unquestioned acceptance as a gold standard for qualitative inquiry. Some argue that the saturation concept is not conceptually consistent with *all* forms of quality inquiry (e.g. O'Reilly & Parker, 2012; Sim et al., 2018b): for Nelson (2016: 5), for instance, "it is not an 'atheoretical' generic research tool that can be applied in any qualitative research design". Low (2019: 131) went further, arguing that saturation defined as no new information "is a logical fallacy, as there are always new theoretical insights to be made as long as data continues to be collected and analysed." We concur with such critique.

However, such critique sits surrounded – often smothered – by the wider conceptualisation of data saturation *as* the gold standard, relatively easily achieved in TA research, a routine item on quality checklists, and championed by various TA proponents and qualitative research titans. Indeed, we hear from researchers who use our *reflexive* TA approach (e.g. Braun & Clarke, 2006; Braun et al., 2019) but reference data or thematic saturation in their publications, because reviewers and editors *required* it, often citing checklists like COREQ or CASP. And researchers often pragmatically acquiesce to reviewers' and editors' demands, even though they hold some critique or question of (data) saturation. For these researchers, the concept of (data) saturation is deployed as the rhetorical device we mentioned earlier, a 'quality assurance' mechanism to get 'passed' by the gatekeepers of knowledge. That

quality checklist criteria can become hoops for researchers to jump through, and actually encourage what many would consider to be bad practice – rather than "improv[ing] the quality of reporting of qualitative research" (p. 356), as the authors of the COREQ checklist hoped – is well acknowledged (e.g. O'Reilly & Parker, 2012; Reicher, 2000).[2]

Where does this leave the TA researcher? Is data saturation a valid or ideal measure for TA sample-size rationalisation? Does demonstrating, or even just claiming, data saturation give validity to the sample sized utilised? Or is data saturation at best unhelpful or meaningless, and at worst problematic, as a concept for sample size in TA? Some clearly see it like that! When Victoria tweeted about writing a commentary entitled "Is saturation a useful concept for TA?" and joked all she had written so far was "No", the tweet garnered numerous virtual high-fives. But others responded with curiosity, asking a version of 'if not saturation, then what?,' demonstrating now much saturation has permeated our qualitative logics. Our answer to these trick(y) questions is – of course – *it depends*. Whether data saturation is a useful concept for TA research depends on how TA, and qualitative researching, are conceptualised, and how data saturation itself is defined and determined. And even when these latter are clarified, the usefulness of data saturation for *reflexive* TA, specifically, is still questionable. Reviewers and editors wielding copies of Guest et al. (2006) or the COREQ checklist, take note: *data saturation is not a universally useful or meaningful concept for all types of TA research* (see also O'Reilly & Parker, 2012).

Data saturation is not a useful concept for all types of TA: Problems and tensions

The authors of empirical explorations of data (and meaning) saturation and sample size tend to offer caveats that *limit* the transferability of their recommendations. As noted, these are often ignored, and advice taken as a more generalised rule. While such poor citation practice is certainly troubling, we are more troubled by the unacknowledged assumptions around both TA *and* saturation in the original papers, which limit the applicability of saturation guidelines. For example, the authors tend to discuss TA (and qualitative content analysis) as if it is a singular method.[3] A general lack of recognition or acknowledgement of *plurality* of TA as a method in 'data saturation experiments' no doubt informs the misperception that such papers provide guidance relevant to all types of TA, including reflexive TA. So a vital first point in considering data saturation and TA is therefore that TA is *not* a singular method.

We generally (currently) distinguish between three main 'types' of TA, which we term coding reliability, codebook and reflexive (Braun et al., 2019). These clusters are divergent in both procedure and underlying philosophy. Authors of 'data saturation experiments' typically use either codebook or coding reliability versions of TA – approaches to TA which centre on the use of a structured *codebook*, determined prior to data analysis, or on the basis of (some) data familiarisation or *some* early coding. The codebook is then typically applied to the entire dataset, in coding reliability TA, or used to document the occurrence of codes in (some) codebook TA. This process for TA coding is *very* different from the open, fluid, organic, and recursive coding practices we advocate for in reflexive TA. In reflexive TA, codes are never finally fixed. They can evolve, expand, contract, be renamed, split apart into several codes, collapsed together with other codes, and even be abandoned. Coding can and often does become more interpretive and conceptual across an analysis, moving beyond surface and explicit meaning to interrogate implicit (latent) meaning. Such developments and refinements reflect the researcher's deepening engagement with their data and their evolving, situated, reflexive, interpretation of them. They also demonstrate a key point for reflexive TA: codes are conceptual tools in the developing analysis and should not be reified into ontologically real *things*. Some of the 'data saturation experiments' discuss code refinement, but it seems to centre on the code definition and inclusion/exclusion criteria, not the nature or scope of the code itself (e.g. Guest et al., 2006). Ando et al. (2014) modified our approach precisely *because* of our lack of a fixed codebook – which they argued was necessary to facilitate the measurement and documentation of data saturation. This in itself suggests an incompatibility between data saturation and an organic reflexive TA approach.

Aspects of TA affecting 'data saturation'

To consider data saturation in and for coding reliability TA in more detail, we return to Guest et al. (2006), who described their analytic approach as follows. An initial codebook was developed for data analysis, including brief and full definitions of codes, guidance on when to, and not to, apply the code, and quotations from the data that provide illustrative examples of the code. The basis on which the codebook was developed is unclear (prior to, or following, some engagement with the data?). The codebook was then applied to the data by two researchers, inter-coder agreement assessed and any discrepancies discussed and

resolved by the research team. The codebook was then revised, and the data recoded by two researchers and inter-coder agreement re-assessed (providing a Kappa score of 0.82, above the 0.8 generally agreed to indicate reliable coding, Yardley, 2008). Themes were identified on the basis of frequency using AnSWR computer software. Analysis of 30 interviews generated 109 content-driven (presumably inductive) codes. The importance of a code was determined by the proportion of interviews to which the code was applied (see also Hennink et al. [2019] who defined high frequency codes in the same way). Thus, there was an emphasis on frequency in determining themes, and data-item frequency in determining the significance of a code. While we do not completely discount the role of recurrence in 'themeyness', we argue that it is only part of what shapes a theme, and the significance of a theme (see also Sim et al., 2018a). Equally, if not more, important is the relevance of the theme to the research question and the *quality* of the theme (Braun & Clarke, 2006; 2012) – does it tell a compelling, coherent and useful story in relation to the research question? Does it offer useful insights that speak to the topic in relation to context and sample?

Different approaches to TA deploy the method in different ways, which affects the potential relevance of data saturation. From the limited information provided, the coding reliability and codebook approaches used in the sample 'data saturation experiments' often rely on a more structured approach to data *collection* than we would advocate for, with reflexive TA. Similar questions need to be asked of participants in interviews, "otherwise, one could never achieve data saturation; it would be a moving target, as new responses are given to newly introduced questions. For this reason, our findings would not be applicable to unstructured and highly exploratory interview techniques" (Guest et al., 2006: 75). Guest et al. (2017) distinguished their method from an inductive qualitative approach, and noted that once piloted, and to facilitate the accurate determination of data saturation, their focus group schedule did not change. The (one) moderator "followed the instrument structure consistently and probed responses to questions, but she did not introduce any information learned in previous focus groups as one typically would in inductive qualitative research" (p. 9).

Such sample size experiments also often use a broadly deductive or 'top down' approach – some or all of the themes are developed ahead of analysis (sometimes from the interview or

focus group guide), or the codebook is developed from analysing the first few interviews and then applied to entire dataset. It is far more difficult, if not impossible, to predict the 'data saturation point' in advance when the analysis is inductive (or deductive in the sense we use it in reflexive TA[4]). And this often connects to the process around data collection. For us *quality* interview data, for instance, are typically 'messy', produced in a context where the interviewer is *responsive* to the participant's developing account, rather than adhering strictly to a pre-determined interview guide (Braun & Clarke, 2013).

In coding reliability and some codebook TA, coding is typically conceptualised as a process of allocating data to pre-determined themes, rather than themes being developed *from* codes, as they are in reflexive TA (Guest et al. [2006] are unusual in identifying themes *from* codes). For example, Hagaman and Wutisch (2017) described the first step of their analytic process as theme identification. Code definitions were then created for the (most common) themes. There can also be slipperiness around the terms *code* and *theme*; these terms, along with the concepts of code saturation and thematic saturation, are often used interchangeably in coding reliability and codebook TA, in contrast to the clear (but not absolute) distinction between codes and themes we see as important in reflexive TA. In reflexive TA, codes and themes represent different levels of complexity: codes capture analytic observations with usually just one idea or facet; themes, constructed *from* codes, are like multi-faceted crystals – they have a core, an 'essence,' which is evident through different facets, each presenting a different rendering of the 'essence'. While staying 'close to' the data, themes in reflexive TA often reflect patterns at both a broader, and more 'abstracted' level than codes, and – even if deductive – are usually difficult to identify in advance of deep analytic work.

Although several approaches to TA acknowledge different *types* of code – such as semantic (surface, obvious, explicit meaning) or latent (implicit, underlying meaning) (Boyatzis, 1998; Braun & Clarke, 2006) – it is rare for 'data saturation experiments' to discuss different types of code and what this might mean, conceptually and practically, in terms of data saturation. Hennink et al. (2017, 2019), with their distinction between inductive and deductive, and concrete and conceptual codes, provide one exception. However, their understanding of what constitutes a conceptual code, on the basis of the examples they present, seems closer to what we would still call semantic codes in reflexive TA, rather than latent (conceptual)

codes as we conceptualise them. For Hennink et al., concrete codes captured "explicit, definitive issues in data; for example, the code 'food taste' captured concrete discussion about the taste of food" (2019: 1486). In contrast, conceptual codes captured abstract constructs – they listed "perceptions, emotions, judgements, or feelings" as examples (p. 1486). They described that "the conceptual code 'denial' captured comments about failure to recognize symptoms of diabetes, refusing testing, or rejecting a diagnosis of diabetes, for example, 'They just don't want to admit that, okay, we have this disease.'" (p. 1486). From *our* perspective, this code "denial" still represents a fairly semantic reading of this extract, based on explicitly-stated content. Similarly, the examples presented from Ando et al. (2014: 5) of higher-level concepts included "remedies for symptoms" and "effect of relapse". Again, these seem to capture a still-semantic reading of data. The code examples in these papers are, then, mostly what we would term descriptive or semantic. This suggests either very 'concrete' data, or a fairly surface-level engagement with the data, and perhaps limited interpretative engagement (Saunders et al., 2017). Morse's (1997) criticism of a coding approach that prioritises consistency and consensus over situated, reflexive interpretation is relevant here. She argued such an approach risks superficiality: "it will simplify the research to such an extent that all of the richness attained from insight will be lost" (Morse, 1997: 446). Data 'saturation' might be facilitated in these approaches, but how is the analysis, interpretation and the potential for new insight potentially foreclosed?

Claims of achieving 'data saturation' in relatively small numbers of interviews or focus groups is likely facilitated not only by the use of semantic focus in coding, but also coding at a relatively coarse level of detail. As an example of this, from research on the health-seeking behaviour of African American men, Guest et al. (2017: 12) presented the example of a code labelled "experimentation." They briefly defined this as "experimentation or research on patients as part of health care"; the full definition directed coders to use the code for "mention or discussion of past or current experiences or beliefs about experimentation" including references to "research studies, guinea pigs, and teaching hospitals, whether actual or perceived". With the acknowledgement that determining the character of a code is partly a contextual judgement – context we do not have access to – this code seems to capture meaning at both the semantic and fairly broad or coarse levels.

The way a theme is conceptualised can also dramatically impact the likelihood that 'data saturation' can be identified (early on). Not all of the papers discussed provide examples of themes. Of those that do, themes tend to be conceptualised as topic-summaries, by which we mean summaries of the range of things participants said, often at an explicit level, in relation to a particular topic or interview/focus group question. This is very different from how themes are conceptualised in reflexive TA – as patterns of shared meaning united by a central concept, developing *out of* the analytic process following coding (Braun & Clarke, 2013, 2019; Braun et al., 2014). But it does fit with the way themes are often conceptualised in coding reliability and codebook versions of TA (see Braun & Clarke, 2019; Braun et al., 2019). For example, one of Ando et al.'s (2014: 5) example themes was titled "impact of MS". In Namey et al. (2016: 437), the themes/codes included "cleanliness of facilities" and "forgetfulness". With themes effectively conceptualised as analytic *inputs*, developed early in, or prior to, the analysis, and/or as topic summaries (perhaps drawn from the interview guide), it seems likely to us that subsequent data collection may contribute additional codes to a theme (e.g. further instances of the "impact of MS"), but that possible or actual themes *will* 'saturate' early. And indeed, if questions asked are used as the basis for subsequent themes, there is a circularity to the analytic process that makes 'data saturation' virtually inevitable.

Different version of TA: Implications for considering (and rejecting) data saturation

There is an important-to-recognise clash of research values that underlie coding reliability and reflexive versions of TA. Coding reliability TA seems to be a firmly neo-positivist activity, prioritising notions of reliability and objectivity of observation valued by positivist quantitative paradigms. Boyatzis (1998), one of the key early authors on TA, presented his ('coding reliability') approach as one that could 'bridge the divide' between the values of positivist (quantitative) and interpretative (qualitative) researchers, but it seems to us more neo-positivist than interpretative-qualitative. By contrast, we expressly developed TA as an approach embedded within, and reflecting the values and sensibility of, a qualitative paradigm; we now call it *reflexive* TA to emphasise this, and to clearly differentiate it (Braun & Clarke, 2019). From our *qualitative* perspective, quality of coding is not demonstrated by 'objective' agreement; coding reliability measures at best demonstrate that coders have been trained to code in the same way using (often coarse and semantic) codes designed to

facilitate the measurement of coding agreement (Yardley, 2008). Coding quality in reflexive TA stems not from consensus between coders, but from depth of engagement with the data, and situated, reflexive interpretation. And this process-based, and organic, evolving orientation to coding makes saturation (especially conceptualised as information redundancy) difficult to align.

For researchers to claim the data were saturated, meaning seems to need to reside *in* data. And sometimes this meaning is treated as fairly self-evident. The data may not even need analysing, with the researcher's impressions of the data during data collection sometimes providing enough of a basis to determine if data saturation has been achieved – an impoverished view of the potential of qualitative research and indeed TA. This conceptualisation of meaning positions the researcher as an archaeologist, excavating meaning from data. Data, code or thematic saturation are possible because there is an imagined concrete basis for determining 'nothing new' to be sought/found. Such an understanding seems to rely on a straightforward realist ontology (Sim et al., 2018a), which we argue is incompatible with the assumptions of reflexive TA. Despite this, as Nelson (2016) noted, the 'information redundancy' saturation concept is invoked even by researchers who subscribe to non-realist ontologies.

Our approach to TA is founded on an entirely different assumptions around meaning – that meaning is not inherent or self-evident in data, that meaning resides at the intersection of the data and the researcher's contextual and theoretically embedded interpretative practices – in short, that meaning requires *interpretation*. On this basis, new meanings are always (theoretically) possible (Low, 2019; Sim et al., 2018a). When we conceptualise research as a situated, reflexive and theoretically embedded practice of knowledge *generation* or *construction*, rather than discovery, there is always the potential for new understandings or insights (Mason, 2010). If we are working with rich, complex, 'messy' data, it will hopefully burst with potential. The challenge will be choosing *what* to explore. We have become infamous for admonishing that 'themes do not emerge' (Braun & Clarke, 2006) – this is not our idea, but we have argued vocally that it is the only way to conceptualise themes for reflexive TA (Braun & Clarke, 2016). From our perspective, attempting to predict the point of data saturation cannot be straightforwardly tied to the number of interviews (or focus groups) in which the theme is evident, as the *meaning* and

indeed meaningfulness *of* any theme derives from the dataset, and the interpretative process. Furthermore, themes are not entities that exist in isolation from one another, themes are chapters in a broader story, and have meaning in relation to other themes (Kerr et al., 2010; Sim et al., 2018a). Codes and coding are likewise context dependent, and particular instances of codes derive at least in part from the particular context in which they are expressed (see Sim et al., 2018a).

Furthermore, in this reflexive organic process, analysis can never be complete (Low, 2019). Coding and deeper analysis don't inevitably reach a fixed end point – instead, the researcher makes a situated, interpretative judgement about when to stop coding and move to theme generation, and when to stop theme generation and mapping thematic relationships to finalise the written report. They can also move back and forth recursively between coding and theme development. So, if reflexive TA researchers use the popular concept of data saturation, the notion of 'no new' makes little sense. But that isn't the only possibly way saturation can be explored or imagined. Akin to Low's (2019) re-conceptualisation of theoretical saturation in grounded theory as pragmatic saturation, what might constitute 'saturation' for reflexive TA researchers is an interpretative judgement related to the purpose and goals of the analysis. It is nigh on impossible to define what will count as saturation in advance of analysis, because we do not know what our analysis will be until we do it. This aligns with Sim et al.'s (2018a) claim that determining sample size in advance is inherently problematic in more interpretative forms of qualitative research. Malterud et al.'s (2016) concept of *information power* – the more relevant information a sample holds, the fewer participants are needed – seems to offer a useful alternative to data saturation for thinking around justifications for sample size in reflexive TA, both actually and pragmatically. The name is seductively concrete enough for the positivist-inclined gatekeeper, the practice flexible enough for qualitative researchers who have fully divested their research practice of positivism (though for a critical discussion of information power, see Sim et al., 2018a).

Beyond data saturation: Sampling as pragmatic practice (as much as anything)

For many, qualitative sample size needs not just an explanation, but some warranty of acceptability. We detect the lingering presence of positivism around discussions of sample size in TA (Vasileiou et al., 2017) – large or probabilistic is best (Guest et al., 2006) – and a

sense of lingering positivist-empiricist produced anxiety. If the sample is not 'reassuringly' large or probabilistic, what criteria could we deploy to justify the adequacy of the sample? Data saturation! As we previously noted, we suspect the concept of data saturation is often deployed as post-hoc rationale or acceptable rhetorical justification of a more pragmatically determined sample size. Data saturation is the rabbit pulled out of the hat, the magic trick that reveals and maybe also conceals.

So, if not data saturation, then what? Determining sample size in qualitative projects is, we suspect, often a pragmatic exercise – not disconnected from what is acceptable or normative: in the local context; in the discipline; to the reviewers and editor of a particular journal; to a particular funding body; to external examiners for a thesis, within the time or financial constraints of a project; and many other factors separate from research design or analytic method… Sample size *can* be determined by a researcher's perception of what research 'gatekeepers' will deem acceptable – and things like editor guidelines which set expected or minimum sample sizes feeds this practice. Experienced qualitative researchers may have developed their own 'rules of thumb' for sample size (Malterud et al., 2016), based on their own expertise, but likely also at least partly informed by such pragmatic considerations. We certainly have our own rules of thumb *and* make pragmatic decisions around sampling.

Is the pragmatic nature of how we might sample for qualitative research a cause for concern? We think it is important to recognise research as nearly always a pragmatic activity, shaped and constrained by the time and resources available to the researcher (Green & Thorogood, 2004; O'Reilly & Parker, 2012), as much as it is also shaped by other things. An 'anxiety' around, perhaps obsession with, qualitative sample size in some quarters is not something that resonates for us – we are comfortable dwelling in a qualitative landscape in which determining sample size relies on a mix of interpretative, situated and pragmatic judgement (Sandleowski, 1995; Sim et al., 2018a).

However, there is often a practical need to determine sample size in advance – for a research proposal, ethics or funding application. In such circumstances, we suggest reflexive TA researchers reflect on the following intersecting aspects of their research: the breadth and focus of the research question; the methods and modes of data collection to be used; identity-based diversity within the population or the desired diversity of the sample; likely

experiential or perspectival diversity in the data; the demands placed on participants; the depth of data likely generated from each participant or data item; the expectations of the local context including discipline; the scope and purpose of the project; the pragmatic constraints of the project; and the analytic goals and purpose of reflexive TA. We suggest then guestimating a *provisional*, anticipated lower and/or upper sample size or range that will potentially generate adequate data to tell a rich, complex and multi-faceted story about patternings related to the phenomena of interest (Sim et al., 2018a). Researchers should then make an in-situ decision about the final sample size, shaped by the adequacy (richness, complexity) of the data for addressing the research question (but with a pragmatic 'nod' to sample size acceptability to the relevant research gatekeepers). Such decisions could and should be made within the process of data collection, reviewing data quality during the process – and recognising that sample size alone is not the only factor at play. Getting *different* stories can require sampling more widely.

Whither data saturation and TA?

Our point here is not that data saturation is never valid and never a useful concept. It might well be – for some forms of TA, and in some circumstances. We can imagine if data collection is underpinned by a realist ontology, follows a fairly structured approach and questions focus on relatively surface-level concerns, data are relatively concrete, participants are relatively homogenous and recruited from a particular setting, and coding focuses on fairly superficial or obvious meaning, with codes as containers for fairly broad topics (e.g. "exercises barriers" and "mood" in Hennink et al., 2019: 1493), then judgements of 'no new' might seem warranted. But data saturation is not the only (valid or invalid) – or indeed the best – rationale for sample size (in TA research). And for reflexive TA, data saturation is an awkward if at times convenient bedfellow, though one perhaps best avoided.

But we know that authors will continue to be asked to explain whether, when, and how data saturation was reached, or the sample size was determined. And that definite answers to questions of TA, sample size and data saturation will continue to be sought. So in the interests of an enriched, more conceptually coherent, and precise or delimited conversation, we encourage authors of any future data (and meaning) saturation 'experiments' to define or provide the following:

- Their conceptualisation of saturation.
- The type of TA they used for the experiment – our typology of coding reliability, codebook and reflexive TA (Braun et al., 2019) is *one* way to differentiate.
- The paradigmatic, ontological and epistemological assumptions in their research.
- Their definitions of a code and a theme, including:
    - Their criteria for determining what constitutes a theme
    - Examples from their codebook, if used
    - Examples of codes and themes

  Readers can then judge for themselves if they share the authors' understanding of what constitutes a code and a theme, and particular types of code and theme (e.g. a concrete versus a conceptual code).
- Justifications of any numerical criteria used in the experiment (e.g. why 10+3 as the stopping criteria, why 3 instances of a theme to determine thematic saturation?).

Providing such information will help readers to determine if they share the authors' paradigmatic and epistemological assumptions about meaningful knowledge and knowledge production, and whether they can safely 'transfer' the guidance around 'how many' to their own use of TA, in their particular context. It would also provide the wider qualitative research community with a better set of tools to question both assertions about (the need for) data saturation (in TA), and the basis on which such assertions are made. Although we have our definite preferences and embedded values for qualitative researching, we are not promoting a singular or narrow take here.

Conclusion

We hope this paper has demonstrated that the same term or concept – here: saturation, code, theme – can have very different *meanings*, and they can be deployed in quite different ways, even within what is ostensibly the same method (TA). This highlights the need for care and reflexivity in describing – and doing – TA (Braun & Clarke, 2019), and in thinking about what elements are at play when evaluating whether saturation (*whatever* that is) is considered for sample size justification.

To address the question posed in the title of the paper: to saturate or not to saturate? We hope our answer – it depends, of course, but often no – is clear by this point. Data

saturation *is* a concept generally coherent for broadly realist, discovery-oriented (coding reliability or codebook) types of TA. However, even there, more precision is needed in how the data saturation concept is defined and determined in discovery-oriented TA research, including in saturation experiments aiming to provide concrete guidance on determining the likely point of data saturation in advance of data collection. But when it comes to reflexive TA, data saturation is *not* a particularly useful, or indeed theoretically coherent, concept.[5] Other concepts – like information power – can offer a more useful way of thinking through data samples. But we recognise that data saturation might be a concept reflexive TA researchers pragmatically chose to deploy to appease research gatekeepers, or might be required to. In doing so, they (and indeed we) are, however, complicit in perpetuating the myth of data saturation as a vital rationale and practice for qualitative research more generally. If a claim of data saturation *must* be deployed for reflexive TA to 'pass go', we encourage researchers to critically comment on this, or provide some justification for it. Or, indeed, perhaps to re-theorise data saturation in new, exciting, and currently unanticipated ways.

Notes

[1] In a parallel focus group study, Hennink et al. (2019) reported that four focus groups were sufficient for code saturation (94% of all codes and 96% of high prevalence codes were identified). However, *meaning* saturation (fully understanding the issues identified through code saturation) required five or more groups. Again, this is not dissimilar to the average number of groups across focus group research (e.g. a mean of 8.4 and median of 5 groups identified by Carlsen & Glenton, 2011). Previously, Guest et al. (2017) had reported that 80% of themes were discoverable in very few (2-3) focus groups, and 90% in 3-6, and claimed three focus groups were enough to identify all of the most prevalent themes. Some have compared (data) saturation in TA from interview and focus group data collection. Namey et al. (2016) reported that eight interviews or three focus groups were necessary to achieve 80% thematic saturation (i.e. 80% of the total number of codes identified) and 16 interviews or five focus groups to achieve 90%. To adequately address a research question focused on evaluation, they recommend sample size between 8 and 16 interviews or three and five

focus groups. An earlier study had identified five focus groups and nine interviews as the point at which (data) saturation was reached (Coenen et al., 2012).

[2] An important wider implication – raised by an anonymous reviewer – is how the inclusion of saturation, and the positioning of saturation as a (required) measure of quality, in these guidelines, might have implications that do not just affect the judged quality and publishability of an individual study. In a context where systematic review and methods like qualitative synthesis deploy 'quality controls' for inclusion, the ramifications are far broader than the individual study, with impacts on what qualitative 'evidence' gets seen and heard through such (highly regarded) mechanisms for assessing evidence for developing, for instance, policy, evidence-based practice, and so forth. We do not have scope to do this point justice here but raise it as a wider quality consideration to be addressed.

[3] Ando et al. (2014) are an exception; they describe their method as a modified *version* of our approach (Braun & Clarke, 2006), involving the addition of a second stage of coding clarifying the initial coding, and the review of codes rather than themes for the purpose of creating a codebook. Yet even so, in claiming that 12 interviews "should be a sufficient sample size for thematic analysis" (p. 7), they nonetheless evoke a singular method of 'thematic analysis'.

[4] The understanding of a 'deductive' approach in coding reliability and codebook TA is often rather different from our conceptualisation – of using existing theory as a lens through which to code and interpret the data. In reflexive TA, using interview questions as themes does not represent a deductive approach just an under-developed analysis (Braun & Clarke, 2006).

[5] Theoretical saturation – whether interpreted as implying a fixed point or not – requires concurrent process of data collection and analysis, and crucially theoretical sampling, practices fairly particular to grounded theory, and not typically elements of a TA.

References

Ando, H., Cousins, R. & Young, C. (2014). Achieving saturation in thematic analysis: Development and refinement of a codebook. *Comprehensive Psychology*, *3*(4), https://doi.org/10.2466/03.CP.3.4.

Bowen, G.A. (2008). Naturalistic inquiry and the saturation concept: A research note. *Qualitative Research*, *8*(1), 137-152.

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks, CA: Sage.

Bragaru, M., van Wilgen, C.P., Geertzen, J.H.B., Ruijs, S.G.J.B., Dijkstra, P.U. et al., (2013). Barriers and facilitators of participation in sports: A qualitative study on Dutch individuals with lower limb amputation. *PLoS ONE*, *8*(3): e59881, https://doi.org/10.1371/journal.pone.0059881

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77.

Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper (Ed.), *APA Handbook of Research Methods in Psychology* (Vol. 2: Research Designs, pp. 57-71). Washington, DC: APA books.

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners.* London: Sage.

Braun, V. & Clarke, V. (2016). (Mis)conceptualising themes, thematic analysis, and other problems with Fugard and Potts' (2015) sample-size tool for thematic analysis. *International Journal of Social Research Methodology*, *19*(6), 739-743.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health, 11*(4), 589-597.

Braun, V., Clarke, V. & Rance, N. (2014). How to use thematic analysis with interview data. In Vossler, A. & Moller, N. (Eds.), *The Counselling & Psychotherapy Research Handbook* (pp. 183-197). London: Sage.

Braun, V., Clarke, V., Terry, G & Hayfield N. (2019). Thematic analysis. In Liamputtong, P. (Ed.), *Handbook of research methods in health and social sciences* (pp. 843-860). Singapore: Springer.

Carlsen, B., Glenton, C. What about N? A methodological study of sample-size reporting in focus group studies. *BMC Medical Research Methodology,* 11, 26. doi:10.1186/1471-2288-11-26

Critical Appraisal Skills Programme (2018). *CASP qualitative checklist*. [online] Available from: https://casp-uk.net/casp-tools-checklists/

Chamberlain, K. (2010). Using grounded theory in health psychology. In M. Murray & K. Chamberlain (Eds.), *Qualitative health psychology: Theory and methods* (pp. 183- 201). London: Sage.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Thousand Oaks, CA: Sage.

Clarke, V., & Braun, V. (2019). Feminist qualitative methods and methodologies in psychology: A review and reflection. *Psychology of Women and Equalities Review, 2* (1), 13-28.

Coenen, M., Stamm, T.A., Stucki, G. & Cieza, A. (2012). Individual interviews and focus groups with patients with rheumatoid arthritis: A comparison of two qualitative methods. *Quality of Life Research*, *21*(2), 359-370.

Constantinou, C.S., Georgiou, M., Perdikogianni, M. (2017). A comparative method for themes saturation (CoMeTS) in qualitative research. *Qualitative Research*, *17*(5), 571-588.

Dey I. (1999). *Grounding grounded theory*. San Francisco, CA: Academic Press.

van Dijk, T. A. (1992). Discourse and the denial of racism. *Discourse & Society, 3*(1), 87-118.

Eynon, M. J., O'Donnell, C., & Williams, L. (2018). Gaining qualitative insight into the subjective experiences of adherers to an exercise referral scheme: A thematic analysis. *Journal of Health Psychology, 23*(11), 1476-1487.

Francis, J.J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M.P. & Grimshaw, J.M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology & Health*, *25*(10), 1229-1245.

Fusch, P.I. & Ness, L.R. (2015). Are we there yet? Data saturation in qualitative research. *The Qualitative Report*, *20*(9), 1408-1416.

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

Green, J. & Thorogood, N. (2004). *Qualitative methods for health research*. London: Sage.

Guest, G., Bunce, A. & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, *18*(1), 59-82.

Guest, G., MacQueen, K., & Namey, E. (2012). *Applied thematic analysis*. Thousand Oaks, CA: Sage.

Guest, G., Namey, E. & McKenna, K. (2016). How many focus groups are enough? Building an evidence base for nonprobability sample sizes. *Field Methods*, *29*,-3-22.

Hagaman, A.K. & Wutich, A. (2017). How many interviews are enough to identify metathemes in multisited and cross-cultural research? Another perspective on Guest, Bunce, and Johnson's (2006) landmark study. *Field Methods*, *29*(1), 23-41.

Hancock, M.E., Amankwaa, L., Revell, M.A. & Mueller, D. (2016). Focus group data saturation: A new approach to data analysis. *The Qualitative Report*, *21*(11), 2124-2130.

Hennink, M.M., Kaiser, B.N. & Marconi, V. C. (2017). Code saturation versus meaning saturation: How many interviews are enough? *Qualitative Health Research, 27*(4), 591-608.

Hennink, M.M., Kaiser, B.N. & Weber, M.B. (2019). What influences saturation? Estimating sample sizes in focus group research. *Qualitative Health Research*, 29(10), 1483-1496.

Kerr, C., Nixon, A. & Wild, D. (2010). Assessing and demonstrating data saturation in qualitative inquiry supporting patient-reported outcomes research. *Expert Review of Pharmaeconomics and Outcomes Research*, *10*, 269-281.

Lincoln, Y.S. & Guba, E.G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Levitt, H.M., Bamberg, M., Creswell, J.W., Frost, D.M., Josselson, R. & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA publications and communications task force report. *American Psychologist*, *73*(1), 26-46.

Low, J. (2019). A pragmatic definition of the concept of theoretical saturation. *Sociological Focus*, *52*(2), 131-139.

Mason, M. (2010). Sample size and saturation in PhD studies using qualitative interviews. *Forum: Qualitative Social Research*, *11*(3).

Malterud, K. (2012). Systematic text condensation: A strategy for qualitative analysis. *Scandinavian Journal of Public Health*, *40*, 795-805.

Malterud, K., Siersma, V.K. & Guassora, A.D. (2016). Sample size in qualitative interview studies: Guided by information power. *Qualitative Health Research*, *26*(13), 1753-1760.

Marshall, A., Donovan-Hall, M., & Ryall, S. (2012). An exploration of athletes' views on their adherence to physiotherapy rehabilitation after sport injury. *Journal of sport rehabilitation, 21*(1), 18-25.

Marshall, J.H., Baker, D.M., Lee, M.J., Jones, G.L., Lobo, A.J. et al. (2018). The assessment of online health videos for surgery in Crohn's disease. *Colorectal Disease*, *20*, 606-613.

Morrow, S.L. (2005). Quality and trustworthiness in qualitative research in counseling psychology. *Journal of Counseling Psychology*, *52*(2), 250-260.

Morse, J.M. (1995). The significance of saturation. *Qualitative Health Research*, *5*(2), 147-149.

Morse, J.M. (2015). "Data were saturated…". *Qualitative Health Research*, *25*(5), 587-588.

Namey, E., Guest, G., McKenna, K. & Chen, M. (2016). Evaluating bang for the buck: A cost-effectiveness comparison between individual interviews and focus groups based on thematic saturation levels. *American Journal of Evaluation*, *37*(3), 425-440.

Nelson, J. (2016). Using conceptual depth criteria: Addressing the challenge of reaching saturation in qualitative research. *Qualitative Research*, *17*(5), 554-570.

O'Reilly, M. & Parker, N. (2012). "Unsatisfactory saturation": A critical exploration of the notion of saturated sample sizes in qualitative research. *Qualitative Research*, *13*(2), 190-197.

Picariello, F., Ali, S., Foubister, C., & Chalder, T. (2017). 'It feels sometimes like my house has burnt down, but I can see the sky': A qualitative study exploring patients' views of cognitive behavioural therapy for chronic fatigue syndrome. *British Journal of Health Psychology, 22*, 383-413.

Reicher, S. (2000). Against methodolatry: Some comments on Elliott, Fischer, and Rennie. *British Journal of Clinical Psychology, 39*(1), 1-6.

Sandelowski, M. (1995). Sample size in qualitative research. *Research in Nursing & Health*, *18*, 179-183.

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H. & Jinks, C. (2017). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity, 52*(4), 1893-1907.

Schweitzer, R., van Wyk, S. & Murray, K. (2015). Therapeutic practice with refugee clients: A qualitative study of therapist experience. *Counselling and Psychotherapy Research*, *15*(2), 109-118.

Sim, J., Saunders, B., Waterfield, J. & Kingstone, T. (2018a). Can sample size in qualitative research be determined apriori? *International Journal of Social Research Methodology*, *21*(5), 619-634.

Sim, J., Saunders, B., Waterfield, J. & Kingstone, T. (2018b). The sample size debate: Response to Norman Blaikie. *International Journal of Social Research Methodology*, *21*(5), 643-646.

Tong, A. Sainsbury, P. & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, *19*(6), 349-357.

Underhill, K., Morrow, K.M., Colleran, C., Holcomb, R., Calabrese, S.K. et al., (2015). A qualitative study of medical mistrust, perceived discrimination, and risk behavior disclosure to clinicians by U.S. male sex workers and other men who have sex with men: Implications for biomedical HIV prevention. *Journal of Urban Health*, *92*(4), 667-686.

Vasileiou, K., Barnett, J., Thorpe, S. & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: Systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology*, 18, 148

Weller, S.C., Vickers, B., Bernard, H.R., Blackburn, A.M., Borgatti S. et al. (2018). Open-ended interview questions and saturation. *PLoS ONE*, *13*(6): e0198606, https://doi.org/10.1371/journal.pone.0198606

Yardley, L. (2008). Demonstrating validity in qualitative psychology. In J. A. Smith (Ed.), *Qualitative psychology: A practical guide to research methods* (pp. 235-251). London: Sage.