Published in final edited form as:

Qual Health Res. 2019 August ; 29(10): 1483–1496. doi:10.1177/1049732318821692.

What Influences Saturation? Estimating Sample Sizes in Focus Group Research

Monique M. Hennink, PhD,

Hubert Department of Global Health, Rollins School of Public Health, Emory University, 1518 Clifton Road, Atlanta, GA 30322, USA

Bonnie N. Kaiser, PhD, MPH, and

Department of Anthropology, University of California, San Diego (previous)

Duke University (current)

Mary Beth Weber, PhD, MPH

Hubert Department of Global Health, Rollins School of Public Health, Emory University

Abstract

Saturation is commonly used to determine sample sizes in qualitative research, yet there is little guidance on what influences saturation. We aimed to assess saturation and identify parameters to estimate sample sizes for focus group studies in advance of data collection. We used two approaches to assess saturation in data from 10 focus group discussions. Four focus groups were sufficient to identify a range of new issues (code saturation), but more groups were needed to fully understand these issues (meaning saturation). Group stratification influenced meaning saturation, whereby one focus group per stratum was needed to identify issues; two groups per stratum provided a more comprehensive understanding of issues, but more groups per stratum provided little additional benefit. We identify six parameters influencing saturation in focus group data: study purpose, type of codes, group stratification, number of groups per stratum, and type and degree of saturation.

Introduction

Selecting an appropriate sample size for qualitative research remains challenging. Since the goal is to select a sample that will yield rich data to understand the phenomenon studied, sample sizes may vary significantly depending on the characteristics of each study. The concept of *saturation* is the most common guiding principle to assess the adequacy of data for a purposive sample (Morse 1995; 2015). Saturation was developed by Glaser & Straus (1967) in their grounded theory approach to qualitative research, which focuses on developing conceptual or explanatory models from textual data. Within this context, *theoretical saturation* (also called theoretical sampling) is used. Theoretical sampling is both continuous and data-driven, involving an iterative process of concurrent sampling, data

corresponding author: mhennin@emory.edu.

The authors declare that there is no conflict of interest.

collection, and data analysis to determine further data sources (Charmaz 2014) and continuing until all constructs of a phenomenon are explored and exhausted to support an emerging theory (Glaser and Strauss 1967). In grounded theory, theoretical saturation therefore focuses on the *adequacy* of a sample to provide conceptual depth and richness to support an emerging theory, rather than on *sample size* per se (Corbin and Strauss 2015; Birks and Mills 2011).

However, the term saturation has become part of the broader lexicon of qualitative research (O'Reilly and Parker 2012). It is now widely used outside of its grounded theory origins and has become "the most frequently touted guarantee of qualitative rigor" (Morse 2015, p.587). In this broader application, it is often called *data saturation* or *thematic saturation* and refers to the point in data collection when issues begin to be repeated and further data collection becomes redundant[i]. This broader use of saturation focuses more directly on assessing sample size rather than the adequacy of data to develop theory (as in theoretical saturation). Yet, it remains unclear what saturation means when used outside of grounded theory, how it can be assessed, and what influences saturation for different qualitative methods, types of data, or research objectives (Nelson 2017; O'Reilly and Parker 2012; Kerr et al 2010). Identifying influences on saturation in this broader context can be used by researchers, reviewers, ethical review boards, and funding agencies to determine effective sample sizes in qualitative research proposals.

Determining the number of focus group discussions needed in a study is a key part of research design, as sample size influences multiple study components (e.g., budget, personnel, timetable). However, saturation (and therefore sample size) cannot be determined in advance, as it requires the review of study data. There is limited methodological research on parameters that influence saturation to assist researchers in selecting an effective sample size in advance of data collection (Morse 1995; Guest *et al* 2006; Kerr *et al* 2010; Carlsen & Glenton 2011; Bryman 2012). Guidance is needed on how to estimate saturation prior to data collection to identify an effective sample size for qualitative research proposals. In this study, we aim to assess saturation in data from focus group discussions and identify parameters to estimate sample sizes for focus group studies in advance of data collection. This goal contributes to continual calls for an evidence base of empirical research on saturation (Morse 1995; Guest *et al* 2006; Kerr *et al* 2010; Carlsen & Glenton 2011).

Assessing Saturation in Focus Group Research

There is a growing concern in qualitative research over researchers claiming to have reached saturation in qualitative studies without providing adequate explanation or justification on how saturation was assessed or achieved (Hennink, Kaiser & Marconi 2016; Malterud et al 2015; Bowen 2008; Guest, Bunce & Johnson 2006; Morse, 1995, 2000, 2015). Carlsen and Glenton (2011) conducted a systematic review of 220 studies using focus group discussions to identify how authors justified their sample size. They found that of those studies that explained the sample size, 83% used saturation to justify the number of focus groups in the study, but few authors stated clearly how saturation was assessed or achieved, and many

[i]This is sometimes called 'information redundancy'.

included unsubstantiated claims of saturation and reported reaching saturation while still using the predetermined number of focus groups. Guest *et al* (2016) similarly reviewed 62 textbooks on qualitative research or focus group methodology and found that 42 provided no guidance at all on the number of focus group discussions needed for a study, 6 recommended saturation, 10 gave a numeric recommendation, and 4 mentioned saturation and provided a suggested number of focus groups. The sample size recommendations ranged widely from 2 to 40 groups, with a commonly cited guideline to conduct at least two focus groups for each defining demographic characteristic. A critical outcome of this review is that none of the recommendations on sample sizes were supported by empirical data demonstrating when saturation is achieved in focus group research. These reviews highlight a significant gap in the methodological literature on empirical research assessing saturation and sample size guidelines for focus group research.

There is a small but emerging body of methodological literature assessing saturation in studies using in-depth interviews (Hennink *et al* 2016; Francis *et al* 2010; Guest *et al,* 2006). However, there are few methodological studies where the authors assess saturation and provide guidance on sample sizes for focus group research. Guest *et al* (2016) used data from a study with 40 focus group discussions to develop empirically based recommendations on sample sizes for focus group studies. In this study, the authors documented the process of identifying codes to ascertain when each code was developed and then determined the number of focus groups needed to identify 80% and 90% of thematic codes across the study. They assessed code frequency across data as a proxy for salience of themes and accounted for any temporal bias in identifying codes by randomizing the order of focus groups and replicating their analyses of saturation. Results showed that 64% of themes were generated from the first focus group, 84% by the third focus group and 90% by the sixth group. This pattern remained regardless of the order in which focus groups were reviewed. Three focus groups were also enough to identify the most prevalent themes across these data. The authors conclude that when averaging the sequential and randomized order of focus groups, two to three focus groups are sufficient to capture 80% of themes, including the most prevalent themes, and three to six groups for 90% of themes in a homogenous study population using a semi-structured discussion guide (Guest *et al* 2016).

In an earlier study, Coenen *et al* (2012) assessed saturation in focus group discussions by different approaches to developing codes: an inductive approach of identifying themes from the data itself and a deductive approach of applying themes to data from an existing theoretical framework. The authors used maximum variation sampling to create a diverse sample of participants, which differs from the largely homogenous sample used by Guest *et al* (2016). Saturation was defined as the point at which linking concepts of two consecutive focus groups revealed no additional second-level categories. The authors deemed that saturation occurred at five focus groups, regardless of the approach to code development.

It is difficult to compare results of these two studies given that they operationalize saturation differently – percentage of codes identified across data (Guest *et al* 2016) and absence of new category development in consecutive focus groups (Coenen *et al* 2012). Nonetheless, across both studies, saturation is achieved by six focus groups. The findings of these studies are significant, as they demonstrate that saturation is achieved at a relatively small number of

focus groups, compared with typical guidance given in methodology textbooks that is not empirically based. Taken together, these studies begin to contribute an understanding of saturation in focus group research using homogenous and diverse samples, amongst inductive and deductively derived codes, and saturation in codes versus categories. Despite the differences in the type of sample, type of codes assessed, and operationalization of saturation, both studies reached saturation at a relatively similar number of focus groups.

However, there are two important limitations of these studies. First, the assessment of saturation is based on identifying the occurrences of new themes, without also assessing the understanding of these themes across the data. Identifying the presence of themes is only the first step in reaching saturation. The first time a theme is identified in data may not be detailed or insightful; therefore, additional data may be required to fully capture the meaning of the issue and to understand the depth, breadth, and nuance of the issue (Kerr *et al* 2010; Hennink *et al* 2016). Thus, the authors of these studies provide no guidance on sample sizes needed to reach saturation in the *meaning* of issues in data. Second, they hardly acknowledge the group format of data collection in focus groups and how this may influence saturation. Focus group discussions involve non-directive interviewing whereby group participants engage in discussion, which generates a different type of data than interviews with a single participant due to the interaction and spontaneity of the group dialogue (Morgan, 1997; Krueger and Casey 2015). The group format has potential to generate a broad range of issues and perspectives, but it may also sacrifice narrative depth and understanding of issues. It is unclear how these elements of focus group discussions influence saturation. Finally, there is no examination of how group composition or demographic stratification of focus groups influence saturation. Assessing how these design elements of focus group research influence sample size and saturation is critically important yet is omitted from current literature.

Study Aims

In this study, we aimed to assess saturation in focus group data and identify parameters to estimate sample sizes for focus group studies in advance of data collection. We utilize the broader application of saturation used outside the grounded theory approach, as described above. This focus is important given that saturation is commonly applied to a wide range of research approaches without adequate description or justification of how it was applied or achieved.

We use two approaches to assess saturation that we developed in an earlier study – *code saturation* and *meaning saturation* (Hennink *et al* 2016) – to assess saturation in focus group data. First, we assessed the sample size needed to reach code saturation, which we define as the point when no additional issues are identified in data and the codebook has stabilized. Second, we assessed the sample size needed to reach meaning saturation, which we define as the point at which we fully understand the issues identified and when no further insights or nuances are found. We then examined code and group characteristics to identify parameters that influence saturation in focus group data. We sought to answer the following specific research questions in this study:

1. How many focus group discussions are needed to reach *code saturation?*

2. How many focus group discussions are needed to reach *meaning saturation?*
3. How do code characteristics and focus group composition influence saturation?
4. What parameters can be used to assess saturation a priori in focus group research?

Methods

Study Background

We used data from the South Asian Health and Prevention Education (SHAPE) study for our analyses (clinicaltrials.gov #NCT01084928). Below we provide an overview of the data collection and analysis of the broader SHAPE study as context for our analyses of saturation in these data. The SHAPE study was a planning and feasibility study to inform the design of a diabetes prevention program for South Asian Americans. South Asians (individuals from the Indian subcontinent) are at a high risk for developing diabetes, often presenting with the condition at younger ages and lower body mass indices than other race-ethnic groups (Gujral et al 2013). Although there is strong evidence from randomized controlled trials showing that lifestyle education interventions can prevent or delay type 2 diabetes in high-risk populations (Crandall et al 2008), there is a need to translate these programs to the South Asian population. SHAPE included a formative phase of qualitative research to inform the development and tailoring of the intervention to the needs of the South Asian community and to ensure its cultural acceptability.

Data Collection and Analysis

SHAPE data comprised focus group discussions with self-identifying South Asians adults living in Atlanta, Georgia. Participants were purposively recruited through advertisements in local South Asian magazines, health fairs, and screenings targeting South Asians, and community locations such as South Asian shopping centers and community organizations. Sixteen focus group discussions were conducted in community locations. Focus groups were stratified by age (18–39 years and 40 years or older) and sex, comprising four groups in each stratum. Focus groups lasted 60–90 minutes and were conducted in English by a trained moderator matched for sex, but not ethnicity, to the participant group. Participants were asked open-ended questions on their views of diabetes and obesity, diet and physical activity behaviors, and barriers and facilitators for a healthy lifestyle, as well as providing feedback on specific design elements of the intervention. Participants were given refreshments, travel reimbursement, and a gift bag. We used data from the first ten focus group discussions in this study, as the final 6 groups focused only on some of the discussion topics and therefore, were not suitable for our analyses of saturation. Data were collected between November 2009 and March 2010. The Emory Institutional Review Board (IRB00019630) approved the study. Individual informed consent was sought from participants before each focus group discussion. Participants were informed of the study procedures, risks and benefits and provided written consent to take part in the focus group discussions and for the audio recording.

All focus groups were digitally recorded, transcribed verbatim, de-identified, and entered into the MaxQDA program (Verbi Software, Germany) for analysis. We conducted a close reading of transcripts to identify issues raised by participants. Each issue was verified by two analysts before its inclusion in a codebook, comprising a codename and a description of each issue. A total of 50 codes were developed including both inductive codes derived from the transcripts and deductive codes originating from the discussion guide. Inter-coder agreement was assessed between two coders to compare the consistency of code use and rectify discrepancies before the whole data set was coded.

To assess saturation, we used a similar process developed in an earlier study (Hennink *et al* 2016). For our analyses on saturation, we documented the process of code development and conducted separate analyses of these procedural data, as described in the sections below.

Assessing Code Saturation

To assess code saturation, we reviewed each focus group discussion transcript in the order in which groups were conducted and documented the development of codes. We recorded all new codes developed and their characteristics, including the code name, code definition, type of code (inductive or deductive), notes about issues with new codes (e.g., clarity of the issue captured, completeness of the code definition), and whether any previously developed codes were present in the transcript. Code definitions included a description of the issue captured, instructions for code application, and an example of text relevant to the code. To document the evolution of code development, we also recorded changes made to codes or code definitions as we proceeded, including the type of change and the focus group at which the change was made. We continued to document code development and the iterative evolution of codes for each focus group discussion until the final codebook was complete.

We then categorized codes for analysis, using the same categorizations as we developed in our earlier work on saturation (Hennink *et al* 2016) as follows. First, codes were categorized as inductive or deductive. Inductive codes were content driven and raised by participants themselves, whereas deductive codes originated from the discussion guide and were then verified with data. Second, changes in code development were categorized as a change in code name and change in code definition (e.g. code expanded, inclusion criteria or examples added). Third, codes were also categorized as concrete or conceptual. ‘Concrete’ codes were those capturing explicit, definitive issues in data; for example, the code ‘food taste’ captured concrete discussion about the taste of food. Similarly, the code ‘family time’ captured any discussion about exercise time competing with family responsibilities. ‘Conceptual’ codes were those capturing abstract constructs such as perceptions, emotions, judgements, or feelings. For example, 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.” These categorizations of codes were used in our analyses to quantify the types of codes developed, types of changes to code development, and timing of code development. Finally, codes were categorized as high or low prevalence. Code prevalence was defined by the number of focus group discussions in which a code was present. On average, codes were present in 7 focus group discussions; therefore, we defined high prevalence codes as those present in more than

7 focus group discussions and low prevalence codes as those present in equal to or fewer than 7 focus groups. In total, there were 27 high-prevalence codes and 23 low-prevalence codes.

To assess whether code saturation was influenced by the order in which focus groups were conducted, we randomized the order of groups and mapped *hypothetical* code development onto the random order. To do this, we randomized the focus groups using a random number generator, but we did not repeat the process of reviewing transcripts given the bias this would have introduced, as this process had already been done with the same transcripts in their actual order. Instead, we assumed that codes would be developed after the same number of repetitions of that issue across the focus groups. For example, in the actual code development, the code 'cultural expectations' was created in focus group 3, after the issue was mentioned in focus groups 1 and 2. Thus, in the random order, we assumed that the same code would likewise be developed after 3 groups in which the issue was raised. Our aim was to reflect researchers' style of code development in the random order as in the actual order, so that we could more directly assess the effect of the order of focus groups on code development. We replicated the pattern of code development in the randomized order of groups by calculating the number of times a code was present (as indicated by the number of focus groups in which the code was applied to the data) before the focus group in which the code was created. We then used these numbers to map hypothetical code development in the randomized order of groups. We then compared hypothetical code development with that from the actual order in which focus groups were conducted.

Assessing Meaning Saturation

We followed the same process to assess meaning saturation (described below) that we used in our previous study on saturation with in-depth interview data (Hennink et al 2016), with the addition of two components to reflect the use of focus group data in this study.

To assess meaning saturation, we selected 19 codes that were central to the aims of the original study on diet, exercise and diabetes and included different types of codes. These codes comprised a mix of concrete (13 codes) and conceptual codes (6 codes) and high prevalence (10 codes) and low prevalence (9 codes) codes (as defined above). This selection reflected the nature of codes developed in this study, whereby there were more concrete than conceptual codes. To assess meaning saturation, we traced these 19 codes to identify what we learned about the code in each successive focus group discussion. This involved using the coded data to search for the code in the first focus group discussion and noting what we learned about this issue from this focus group, then searching for the code in the next focus group and noting any new aspects or nuances of the code from that group, and continuing until all 10 focus groups had been reviewed. This process was repeated for all 19 codes that were traced. For each code, we noted at which focus group there were no new aspects of a code raised and no further understanding of the code, only the repetition of earlier aspects. We deemed this as the point of meaning saturation for that code. We then compared the number of focus group discussions needed to reach meaning saturation with the number needed to reach code saturation determined in our earlier analyses.

To assess whether meaning saturation is influenced by the type of code, we compared the timing of saturation for concrete and conceptual codes. Concrete codes included: ‘family time’, ‘homeopathy’, ‘exercise instructor’, ‘exercise measures’, ‘exercise gender’, ‘exercise venues’, ‘physical appearance’, ‘ingredient cost’, ‘food taste’, ‘diabetes cause’, ‘US-Indian food’, ‘exercise barriers’, and ‘exercise perception’. Conceptual codes included: ‘denial’, ‘exercise pleasure’, ‘work success’, ‘women’s responsibility’, ‘mood’, and ‘cultural expectations’. To assess whether meaning saturation is influenced by the prevalence of a code, we compared saturation by high and low prevalence codes.

To assess whether meaning saturation is influenced by the number of participants who discussed a code, we noted the number of participants contributing to the discussion of each code across all focus groups. If 4 people had discussed a code in the first focus group, 2 in the second, and 6 in the third, we determined that a total of 12 participants had discussed this code across the data. We then identified whether there was any pattern in saturation by the number of participants discussing a code. Finally, to assess how saturation is influenced by the demographic stratification of the focus groups (described earlier), we noted the age and sex composition of each group on the trajectories and identified any patterns in saturation by these strata.

Results

Part I: Code Saturation

Code development—Figure 1 shows the timing of code development across all focus groups in the study. The figure shows the focus group discussions in the order in which they were conducted, the number of new codes developed in each successive focus group, the type of code developed (inductive, deductive), and the demographic stratum of each focus group. A total of 50 codes were developed in the study comprising 58% inductive and 42% deductive codes. Deductive codes were developed only from focus groups 1 and 2, with only inductive codes added thereafter. The vast majority of codes (60%) were identified in the first focus group discussion, with a sharp decline in new codes after this. From the second focus group, an additional 12 codes were developed, with 84% of codes developed at this point. Focus groups 3 to 6 added 8 new codes, with only a few new codes per focus group; most of these new codes (5/8) were of low prevalence across the data. After focus group 6, no further new codes were developed.

Given that the majority of new codes (60%) were identified in the first focus group discussion, we assessed whether the order in which the focus groups were conducted influenced the pattern of new code development and code saturation, particularly given the demographic stratification of the focus groups in this study. To assess this, we compared the pattern of code development in the *actual* order in which focus groups were conducted with a *randomized* order of focus groups. Figure 2 shows the same pattern of code development in both the actual and the randomized order of focus groups, with approximately 60% of codes developed in the first focus group discussion and a strong decline in new codes identified in subsequent focus groups. We also find that both scenarios reach saturation with over 90% of codes developed at focus group 4 (94% and 92% in the actual and random

order, respectively). Therefore, the order in which focus groups are conducted has little influence on the pattern of new code development or on code saturation.

Codebook development—We recorded codebook development by documenting the timing of changes to codes and code definitions (Table 1). Most changes to code definitions occurred in focus groups 2 and 3, with no changes to the codebook occurring after focus group 6. The two most common types of code definition changes were adding examples and conceptually expanding a code definition, which consisted of adding a new dimension of the issue to the code definition. For example, the original code definition of the code 'exercise with friends' was "Discussion about whom to exercise with in the intervention (e.g., friends, other South Asians)." Following focus group 2, this code definition was expanded to include "social support to exercise", and following focus group 5, it was expanded again to include the concept of "group accountability as a motivator to exercise." While over half of the codes (58%) were inductive, most of the code definition changes (84%) were made to deductive codes, to ensure that these codes, which were derived externally from the data, effectively reflected the issue raised in the data. Some codes were refined multiple times, with over one-third of the code definition changes made to only three codes ('exercise perception', 'exercise barriers', and 'healthy diet barriers').

Code prevalence—To identify when more or less prevalent codes were developed, we examined code development by code prevalence and type of code. Figure 3 depicts each code as a separate bar: the location of a code on the x-axis indicates the focus group in which the code was developed, and the height of the bar shows the number of focus groups in which the code was used. For example, the first 13 bars show that these codes were all developed in the first focus group discussion and were all high prevalence codes, present in all 10 focus groups. The dashed line indicates the average number of focus group in which a code was used – about 7 focus groups. Figure 3 shows that 27 codes were of high prevalence (above the line), and 23 were of low prevalence (below the line) across all data. The majority of high-prevalence codes (81%, 22/27) were identified in the first focus group discussion, and by the third focus group, 96% (26/27) of all high-prevalence codes were identified. Thus, the vast majority of high-prevalence codes were identified in early focus group discussions. Most low-prevalence codes (65%, 15/23) were developed after the first focus group, with a clustering at focus group 2. This shows that more focus groups are needed to identify low-prevalence codes.

Figure 3 also shows when the different types of codes (concrete or conceptual) were developed and their prevalence across the data. The first focus group almost exclusively generated concrete codes (97%, 29/30), most of which were also of high prevalence (76%, 22/29). In contrast, half of the conceptual codes were low prevalence (50%, 3/6) and developed later, in focus groups 2, 3, and 6. Overall, codes developed in early focus groups were high prevalence concrete codes, while those developed in later focus groups were mostly low prevalence and included more conceptual codes.

Code saturation—We did not use a set threshold to determine code saturation but were guided by the results of our analyses. We determined that code saturation was reached after four focus group discussions, based on code identification (94% of codes had been

identified), code prevalence (96% of high-prevalence codes were identified) and codebook stability (90% of codebook changes were made). Therefore, four focus group discussions were sufficient to *identify* the range of issue present in these data.

Part II: Meaning Saturation

Having established that code saturation was reached at four focus group discussions, we then explored whether four groups are also sufficient to reach meaning saturation, whereby we gain a comprehensive *understanding* of the issues raised. To assess meaning saturation, we traced a range of codes across all focus groups and noted what we learned about each code from successive focus groups until no more new dimensions of the code were uncovered. Meaning saturation was deemed to be reached at the last focus group discussion in which a new dimension of the code was identified. For example, the code 'food taste' was identified in the first focus group discussion and reached meaning saturation at the second focus group, as no more new aspects of this code were identified (after this point, there was only repetition of dimensions already identified). The code 'work success' was also identified in the first focus group discussion, but new aspects of the code were identified in the fourth and sixth focus group but none thereafter, so this code reached meaning saturation by focus group six. We traced 17 codes to assess meaning saturation, comprising a mix of concrete and conceptual codes, and high- and low-prevalence codes. Below we assess the influence of these factors on meaning saturation of these codes.

Figure 4 shows the results of the code tracing, indicating the focus group at which each code was first identified in data and the focus group at which it reached meaning saturation. Most codes were identified in the first focus group discussion, but they did not reach meaning saturation until later focus groups. This shows that data from *multiple* focus groups are needed to understand many of the issues, with successive focus groups adding different dimensions of a code until a more complete understanding of the issue is reached. Codes also reached meaning saturation at different points in the data, some requiring more data to fully understand the issue. Both concrete and conceptual codes needed data from a range of focus groups to reach meaning saturation. For example, the concrete code 'exercise gender' and the conceptual code 'work success' needed data from 6 and 7 focus groups, respectively, to capture the various perspectives on each of these issues. This shows that reaching saturation needs to go beyond code saturation (whereby codes are *identified in* data) towards meaning saturation to fully *understand* the issues raised, and capture the different dimensions, context, and nuances of the issues.

Figure 4 showed that codes reached meaning saturation at different points in the data. Below, we examine a range of influences on reaching meaning saturation, including the type of code (concrete or conceptual), code prevalence, demographic strata of focus groups, and the number of participants who discussed an issue across data.

We explored when different types of codes (concrete or conceptual) reached meaning saturation. Concrete codes reached meaning saturation at different points. A few concrete codes reached meaning saturation after one or two focus groups, for example the codes 'exercise instructor' and 'food taste' (Figure 4). Several concrete codes clustered to reach meaning saturation at focus group 5. The remaining concrete codes reached meaning

saturation later, at focus groups 6 to 11. Conceptual codes showed a more consistent pattern in reaching meaning saturation. All conceptual codes were first identified in focus group 1 or 2, but they did not reach meaning saturation until focus group 5 or later, with one code not reaching saturation (‘cultural expectations’). No conceptual codes reached meaning saturation in fewer than five focus groups, which shows that more data are needed to reach meaning saturation for conceptual codes than for some concrete codes. Overall, with the exception of three concrete codes, all other codes reached meaning saturation at focus group 5 or later.

We examined how code prevalence (high or low prevalence) influences meaning saturation. Figure 4 shows that high-prevalence codes reached meaning saturation between focus groups 4 and 10, with a clustering at focus groups 5 and 6. High-prevalence conceptual codes needed more focus groups to reach meaning saturation (6–10 focus groups) compared with high-prevalence concrete codes (4–5 focus groups). Low-prevalence codes reached meaning saturation between focus groups 2 and 8 and also showed a clustering at focus groups 5 and 6. While all low-prevalence conceptual codes reached saturation at focus group 5 or 6, low-prevalence concrete codes showed a much more mixed pattern, reaching saturation between focus groups 2 and 8. These results show that even high-prevalence codes require a range of focus groups to fully understand issues, with high-prevalence conceptual codes requiring more data than high prevalence concrete codes.

The strongest pattern in meaning saturation was found by demographic strata of the focus groups. Each focus group was stratified by age (younger/older) and sex (men/women), with two to three focus groups conducted within each stratum across the study. Figure 4 indicates the order in which focus group discussions were conducted and the characteristics of each stratum. At focus group 5, each stratum had been included once, and by focus group 10, each stratum had been included two to three times. Results show that regardless of the type of code or code prevalence (described above), meaning saturation clusters at focus group 5 once all strata have been included once. This shows that for many codes, meaning saturation is reached once the perspectives of all strata have been included, and therefore the diversity within each code has been captured. This is shown in Table 2, which exemplifies the different dimensions of a code that are captured in each demographic stratum. For example, with the concrete code ‘exercise barriers’, the first focus group identified issues specific to this stratum of young men (e.g., priority for education over exercise and physical appearance), in addition to other issues. The second focus group adds the perspective of older women about the lack of awareness of exercise benefits beyond weight loss. The third and fourth focus group add a range of issues related to older men, such as barriers of weather and managing an exercise routine. The fifth focus group adds the perspective of young women, regarding the need to be accompanied to exercise classes and the difficulty of exercising at home. Once the perspectives of all strata are included, this code reached meaning saturation, and thereafter only repeated issues were found in later focus groups. Similarly, the conceptual code ‘mood’, which identified emotions related to diet and exercise, captures novel aspects of the code from each of the different strata before it reached meaning saturation.

There are four codes (in Figure 4) that reached meaning saturation before all strata were completed. These were all concrete codes (‘family time’, ‘food taste’, ‘exercise instructor’ and ‘diabetes cause’), with little variation in the way the issue was discussed across all data. For example, the code ‘food taste’ repeatedly focused on the issue that healthy food was not tasty, with no variation across focus groups, while the code ‘exercise instructor’ highlighted that participants prioritized the experience of an instructor over their cultural background, with no further nuances of this issue across focus groups. These results show that codes that are nuanced by the characteristics of demographic strata (e.g., age or sex) reach saturation only after all strata are included, while codes that are not influenced by demographic strata will reach saturation without all strata included.

Given our results on the influence of demographic strata on meaning saturation, we assessed whether more than one focus group per stratum was needed to reach meaning saturation. Results show that some codes needed data from multiple focus groups per stratum to reach meaning saturation. These included both concrete codes (‘exercise venues’, ‘physical appearance’) and conceptual codes (‘cultural expectations’, ‘work success’). The clustering of codes reaching meaning saturation at focus group 6 also shows that a range of codes needed more than one focus group from the strata with young men and older men to reach meaning saturation. For the strata that included three focus group discussions in each – older men and older women – only one code (‘exercise gender’) identified new elements from all three focus groups with older women, suggesting this issue is nuanced only for older women. Other codes showed no new issues in the third focus group in these strata. These results show that conducting 1–2 focus groups per stratum is likely to contribute to a more comprehensive understanding of a code but suggests limited value in conducting three groups per stratum.

We also examined whether the total number of participants discussing a code influenced meaning saturation but found no clear patterns. Codes discussed by a high number of participants across data (e.g., 51) and those discussed with a low number of participants (e.g., 9) both reached meaning saturation at the same point (focus group 5). More participants discussed high-prevalence codes than low-prevalence codes and concrete codes versus conceptual codes; however, there were no patterns by meaning saturation.

Discussion

Through this study, we contribute to a small but growing evidence base of empirical research on saturation. We sought to assess saturation using two approaches - code saturation and meaning saturation - and to develop parameters of saturation to estimate and justify sample sizes for focus group research in advance of data collection.

Our results show that *code saturation* was reached at four focus group discussions, whereby 94% of all codes and 96% of high-prevalence codes had been identified, and the codebook had stabilized. The first focus group generated 60% of all new codes with a sharp decline thereafter, regardless of the order in which focus groups were conducted. Most codes developed in early focus groups were high-prevalence, concrete codes. Comparing these results with our earlier study on saturation in in-depth interviews (Hennink et al 2016), we

also found the first in-depth interview generated the majority of new codes (53%), most of which were also concrete and high-prevalence codes. While the first focus group generated more new codes than the first in-depth interview, it is not remarkably higher considering issues are generated by a group of participants. These results also reflect the findings of other studies examining saturation in focus group data, whereby code saturation was reached at 5 focus groups (Coenen *et al* 2012) and between 3 to 6 focus groups (Guest *et al* 2016). These collective findings provide important evidence that relatively few focus groups are needed to generate a majority of new issues in a study. This contradicts general guidelines provided in academic literature (albeit not based on empirical research) recommending much higher numbers of focus groups in a study (e.g., 10, 20, or 40 focus groups). However, it is important to remember that code saturation identifies the presence of issues in data, in particular high-prevalence concrete issues, but may not provide a full understanding of all issues, their diversity, or nuances. This goal may be suitable for some research objectives, particularly for designing research instruments or interventions; however, these limitations of code saturation should be borne in mind if using this strategy.

We found that reaching *meaning saturation* requires more data than code saturation. While four focus groups were sufficient to identify the majority of issues across the data, more data were needed to fully understand these issues. Our results showed that even issues identified in the first focus group discussion needed more data to fully understand the issue, regardless of the type of code (concrete/conceptual; high/low prevalence). In addition, codes reached meaning saturation at different points in the data; some codes required much more data than others to reach meaning saturation. Even low-prevalence and conceptual issues contributed to building a comprehensive understanding of a phenomenon; therefore, the prevalence of issues in data does not indicate their significance in understanding the study phenomenon. As indicated by Morse (1995, p148), “it is often the infrequent gem that puts other data into perspective, that becomes the central key to understanding the data…it is the *implicit* that is interesting”. Reaching meaning saturation thus relates to the “informational power” (Malterud *et al* 2015) of the sample to provide depth of understanding of the issues. It goes beyond identifying the presence of issues and moves towards gaining “conceptual depth” to capture the range, complexity, subtlety, resonance, and thereby the validity of issues in data (Nelson 2016). Identifying a sample size adequate to meet these characteristics is critical to maximize the benefits of conducting qualitative research.

The most consistent influence on meaning saturation was the demographic strata of the focus groups. Meaning saturation clustered at focus group five for most concrete codes – this represents the point at which at least one focus group from each demographic stratum was included. This indicates that once the perspectives of each demographic stratum have been captured, meaning saturation was reached on most codes. This finding is compelling because it shows that codes that are nuanced by the characteristics of demographic strata will reach saturation only after all strata are included, while codes not influenced by demographic strata will reach saturation without all strata included, as these issues have less diversity by these characteristics. In our data, the topics diet and exercise are highly nuanced by both sex and age for the South Asian study population; therefore, we continued to identify new insights across codes with each demographic stratum until all strata were included, with few

new insights thereafter. Most conceptual codes needed more data, beyond one focus group per stratum to reach meaning saturation.

Many researchers stratify focus groups by demographic characteristics precisely because they anticipate different nuances to emerge from the various strata and to enable analytic comparisons to distinguish patterns in data. The most useful strata to use are often guided by research literature and built into the study design. Overall, it is not the number of groups per se that determines meaning saturation but the point at which all strata are included in the study – in our study this was at five focus groups, but for other studies this point may be different. A common guideline for focus group research is to conduct *at least* two focus groups for each demographic stratum in the study (Krueger and Casey 2015; Barbour 2007; Fern 2001; Greenbaum 2000; Morgan 1997). Our results support conducting two groups per stratum to provide a more comprehensive understanding of issues in particular to fully capture nuances of conceptual codes. However, we found little additional benefit in conducting more than two groups per stratum.

These results have important implications for estimating sample sizes a priori for focus group studies. Based on our findings, we recommend using both the number of strata and number of groups per stratum as key criteria to identify an adequate sample size to reach meaning saturation. For example, researchers doing a focus group study stratified by one characteristic (e.g., sex) would need to conduct two groups to include both strata but should ideally conduct two groups for each of these strata - thereby making a sample size of 4 focus groups. Researchers doing a study using two strata (e.g., sex and age) would need to conduct four groups to include all strata (e.g. younger women, older women, younger men, and older men) and ideally should conduct two groups for each of these strata, for a total sample size of eight focus groups. While this strategy may not be new to seasoned qualitative researchers, our study provides the empirical evidence that was previously lacking to support this approach and gives clear justification for why more groups per stratum are not necessarily better. For focus group studies where groups are not stratified by demographic characteristics, authors of other empirical studies have provided guidance on reaching saturation - Guest et al (2016) show that a homogenous study population where focus groups are not stratified can reach saturation in three to six focus groups (Guest *et al* 2016), and even with a more diverse study population, saturation may be reached at five focus groups (Coenen *et al* 2012).

We propose that an adequate sample size to reach saturation in focus group research depends on a range of parameters and is likely to differ from one study to the next. Therefore, providing universal sample size recommendations for focus groups studies is not useful. Instead we present a range of parameters based on our study findings that influence saturation in focus groups, which can be used to estimate saturation across different studies (Table 3). These parameters include the study purpose, code characteristics, group composition, and desired type and degree of saturation. Each parameter needs to be considered individually, but the estimated sample size is determined by the combination of all parameters rather than by any single parameter alone. For example, one parameter may suggest a smaller sample size, but collectively they may indicate a larger sample is needed. Therefore, researchers need to assess each study by its specific characteristics and how these

may influence saturation to determine an appropriate sample size. Although saturation is ultimately determined during data collection, these parameters provide guidance on identifying *and justifying* a sample size a priori, such as for a research proposal, but they can equally be used to justify the basis on which saturation was assessed or achieved in a completed study. Often the justifications for sample sizes or reaching saturation are absent in published qualitative research, perhaps because there is little empirical guidance on how to do this. Sample size estimates also need to remain flexible to allow the inductive process to be used during data collection; often this is achieved by identifying a range rather a fixed number when a sample size is proposed in advance (e.g., 4–6 focus groups).

It is also important to remain pragmatic on the degree of saturation sought (e.g., 80% or 90%). While it is near impossible to reach total saturation, since there is always the potential to discover new things in data, this is also not the objective of saturation (Corbin and Straus 2008; Saunders *et al* 2017). It is not reaching a particular benchmark that is critical but reaching a point where it is determined that new discoveries do not add further insights, thus reaching a point of 'diminishing returns' in terms of developing a sufficiently robust understanding of the phenomenon (Mason 2010). While this assessment can only be made during data review, our study provides useful guidelines on when this may occur in focus group data.

Study Limitations

To assess the effect of focus group order on saturation, we used a hypothetical randomization of focus group order, rather than repeating the process of code development using the randomized order of focus group discussions. While the benefits of this approach outweighed the risk of bias had the same researchers repeated the code development process, we could have recruited another group of researchers to conduct this task. Additionally, there is a chance that our findings are influenced by coding preferences and practices of the researchers involved. For example, researcher's coding style (e.g., lumper vs. splitter) could affect the number and scope of codes developed, and other researchers might have had slightly different findings regarding timing of saturation had they taken a very different approach.

Conclusion

With this study, we contribute empirical research to identify influences on saturation in focus group research. We examined two approaches to assessing saturation and use our results to develop parameters of saturation that may be used to determine effective sample sizes for focus group studies in advance of data collection. Our results show that reaching code saturation captures the breadth of issues and requires few focus groups, while achieving meaning saturation requires more focus groups for greater depth and understanding of these issues. We also identify the strong influence of demographic strata of focus groups on saturation and sample size. If saturation continues to be hailed as the criterion for rigor in determining an adequate sample size in qualitative research, still further research is needed to examine the nature of saturation in different types of data, data collection methods, and research approaches.

References

Barbour R (2007). Doing Focus Groups. The SAGE Qualitative Research Kit. Sage Publications: London.

Birks J, and Mills J (2011) Grounded Theory: A Practical Guide. Sage Publications:London.

Bowen G (2008). Naturalistic inquiry and the saturation concept: A research note. Qualitative Research, 8, 137–152. doi:10.1177/1468794107085301.

Bryman A (2012). Social Research Methods. Fourth Edition, Oxford University Press, Oxford UK.

Carlsen B, Glenton C (2011). What about N? A methodological study of sample-size reporting in focus group studies. BMC Medical Research Methodology, 11, Article 26. doi: 10.1186/1471-2288-11-26.

Charmaz K (2014). Conducting Grounded Theory. A Practical Guide through Qualitative Analysis. Second Edition. London; SAGE.

Coenen M, Coenen T, Stamm A, Stucki G, and Cieza A (2012). Individual interviews and focus groups with patients with rheumatoid arthritis: A comparison of two qualitative methods. Quality of Life Research, 21:359–70. doi: 10.1007/s11136-011-9943-2. [PubMed: 21706128]

Corbin J and Strauss A (2008). Basics of Qualitative Research: Techniques and Procedures for developing grounded theory. Sage: Thousand Oaks, CA.

Corbin J and Strauss A (2015). Basics of Qualitative Research: Techniques and Procedures for developing grounded theory. Fourth Edition. Sage: Thousand Oaks, CA.

Crandall JP, Knowler WC, Kahn SE, et al. (2008). The prevention of type 2 diabetes. Nat Clin PractEndocrinolMetab. 2008:4(7):382–393. doi: 10.1038/ncpendmet0843.

Fern E (2001). Advancing Focus Group Research. Sage: Thousand Oaks: CA

Francis J, Johnson M, Robertson C, Glidewell L, Entwistle V, Eccles M, Grimshaw J (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. Psychology & Health, 25, 1229–1245. doi: 10.1080/08870440903194015. [PubMed: 20204937]

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

Greenbaum T (2000). Moderating Focus Groups A Practical Guide for Group Facilitation. Sage: Thousand Oaks: CA

Guest G, Bunce A, Johnson L (2006). How many interviews are enough? An experiment with data saturation and variability. Field Methods, 18, 59–82. doi: 10.1177/1525822X05279903.

Guest G, Namey E, McKenna K, (2016). How many focus groups are enough? Building an Evidence Base for Non-Probability sample sizes. Field Methods, Vol 29(1), 3–22. doi: 10.1177/1525822X16639015.

Gujral UP, Pradeepa R, Weber MB, Narayan KM, Mohan V. (2013). Type 2 diabetes in South Asians: similarities and differences with white Caucasian and other populations. Ann N Y Acad Sci. 2013:1281:51–63. doi: 10.1111/j.1749-6632.2012.06838.x. [PubMed: 23317344]

Hennink M, Kaiser B, and Marconi V (2016). Code saturation versus meaning saturation: How many interviews are enough? Qualitative Health Research, Vol 27(4). doi: 10.1177/1049732316665344.

Kerr C, Nixon A, Wild D (2010). Assessing and demonstrating data saturation in qualitative inquiry supporting patient-reported outcomes research. Expert Review of Pharmacoeconomics & Outcomes Research, 10, 269–281. doi: 10.1586/erp.10.30. [PubMed: 20545592]

Krueger R and Casey M (2015) Focus Groups: A Practical Guide for Applied Research, Fifth edition. Thousand Oaks, CA: Sage Publications.

Malterud K, Siersma V, Guassora A (2016). Sample size in qualitative interview studies: Guided by information power. Qualitative Health Research. Vol 26(13), 1753–1760. doi: 10.1177/1049732315617444. [PubMed: 26613970]

Mason M (2010). Sample size and saturation in PhD studies using qualitative interviews. Forum for Qualitative Health Research, 11(3), art 8. doi: 10.17169/fqs-11.3.1428.

Morgan D (1997) Focus Groups as Qualitative Research, Second edition, Qualitative Research Methods Series, vol. 16 Thousand Oaks, CA: Sage Publications.

Morse J (1995). The significance of saturation [Editorial]. Qualitative Health Research, 5, 147–149. doi: 10.1177/104973239500500201.

Morse J (2000). Determining sample size [Editorial]. Qualitative Health Research, 10, 3–5. doi: 10.1177/104973200129118183.

Morse J (2015). Data were saturated … [Editorial]. Qualitative Health Research, 25, 587–588. doi: 10.1177/1049732315576699. [PubMed: 25829508]

Nelson J (2016). Using conceptual depth criteria: addressing the challenge of reaching saturation in qualitative research. Qualitative Research. 17(5). doi: 10.1177/1468794116679873.

Patton M (2015) Qualitative Research and Evaluation Methods, 4rd edition. Thousand Oaks, CA: Sage Publications.

Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, Burroughs H and Jinks C (2017). Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant, doi: 10.1007/s11135-017-0574-8.

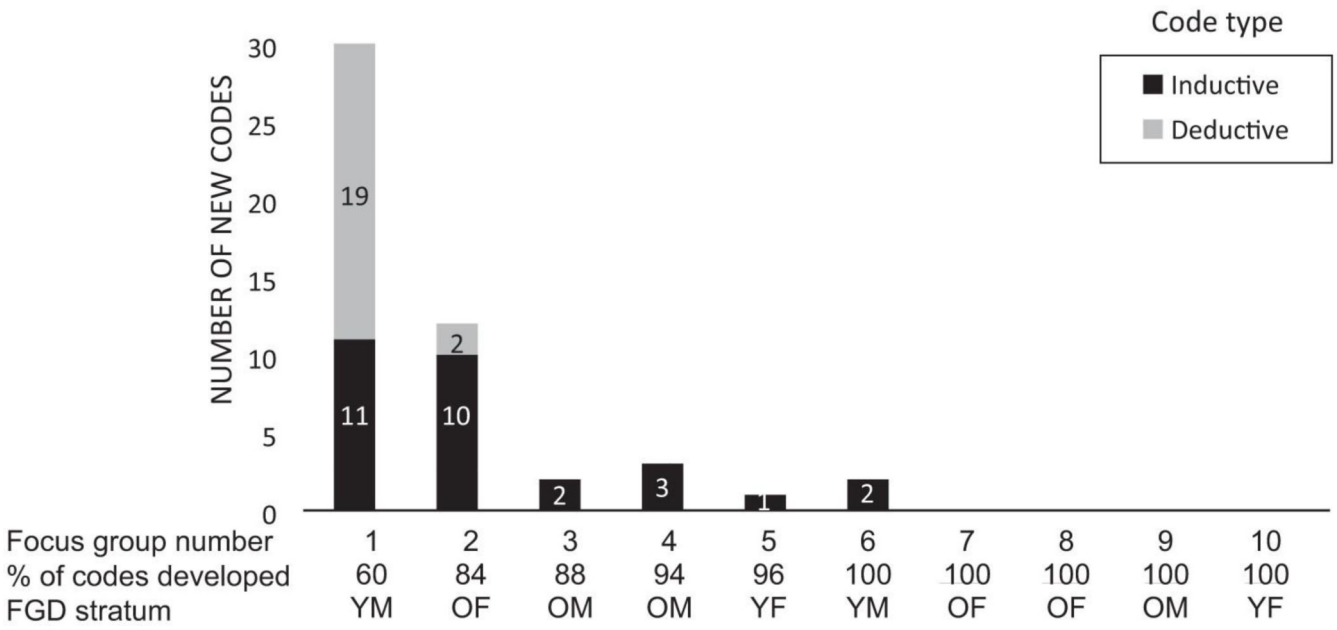

Figure 1.

Timing of code development and code saturation.

*Note.*Two deductive codes were developed in Focus Group 2. These codes were derived from questions in the discussion guide that were not probed in Focus Group 1 but were probed in Focus Group 2. FGD = focus group discussion; Y = Younger; M = Male; O = Older; F = Female.

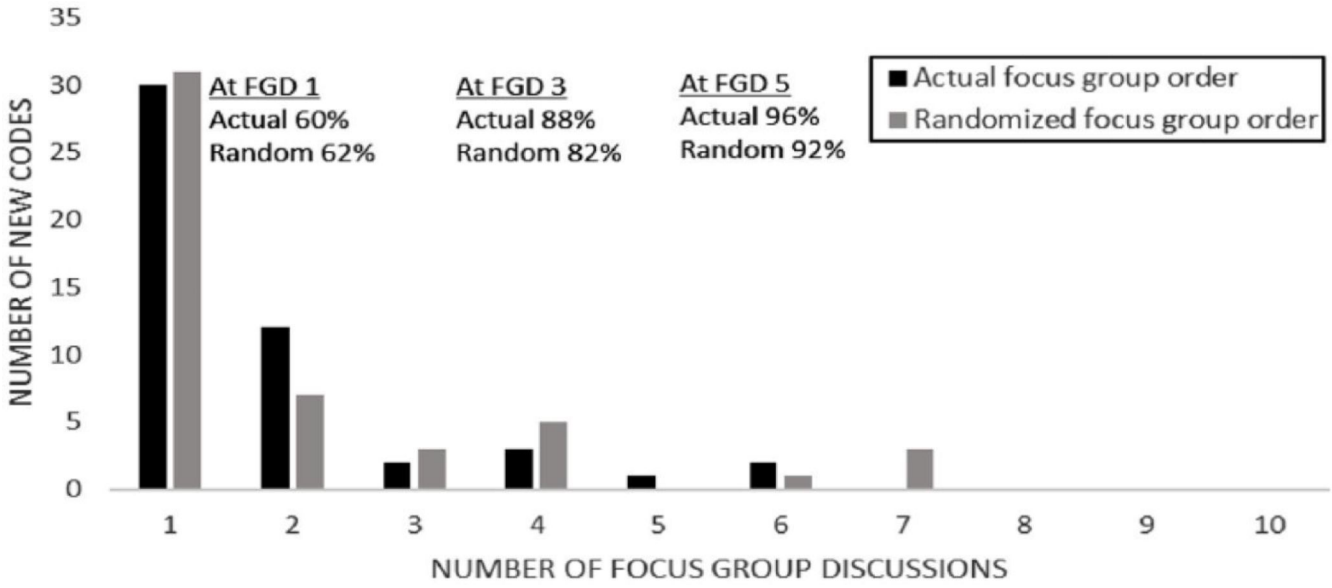

Figure 2.

Timing of code development—Actual versus randomized order of FGDs.

*Note.*FGD = focus group discussion.

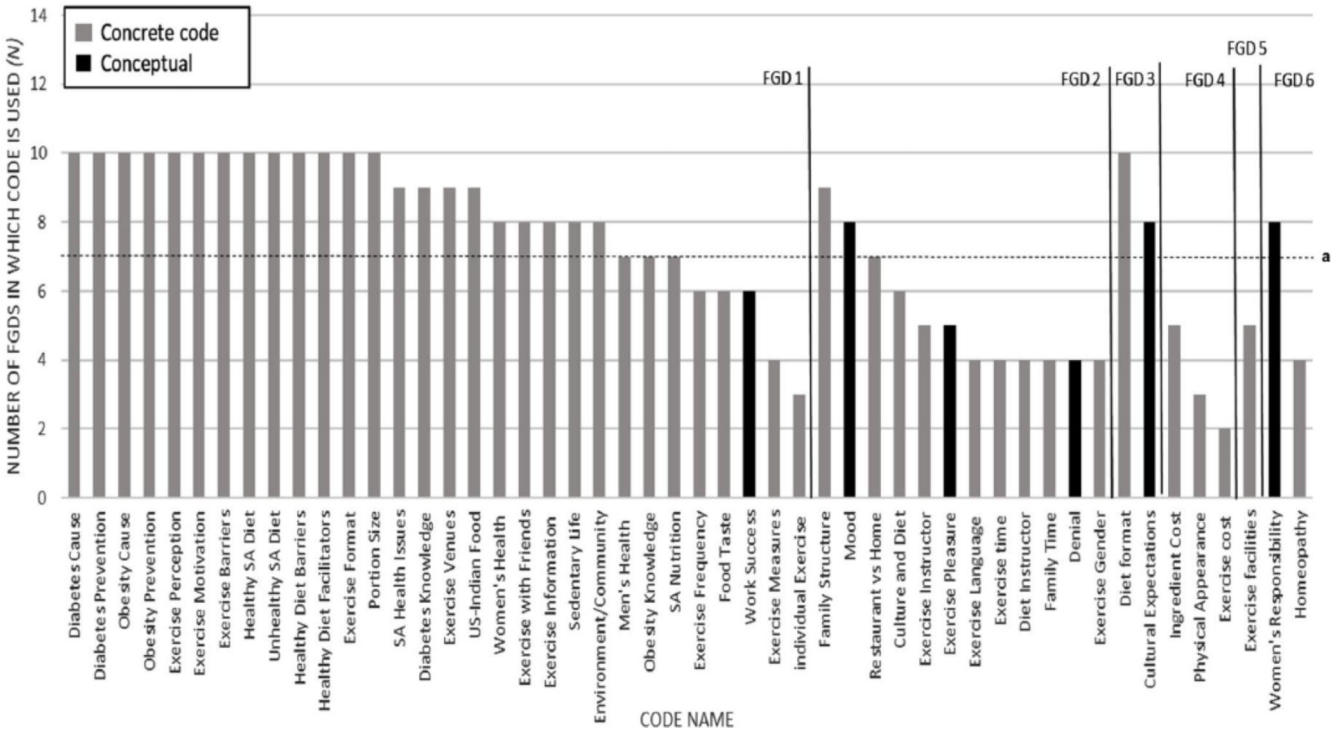

Figure 3.

Timing of code development by code prevalence and type.

Note. FGD = focus group discussion.

[a]Dashedline indicates average code prevalence across all data at 7.2 FGDs.

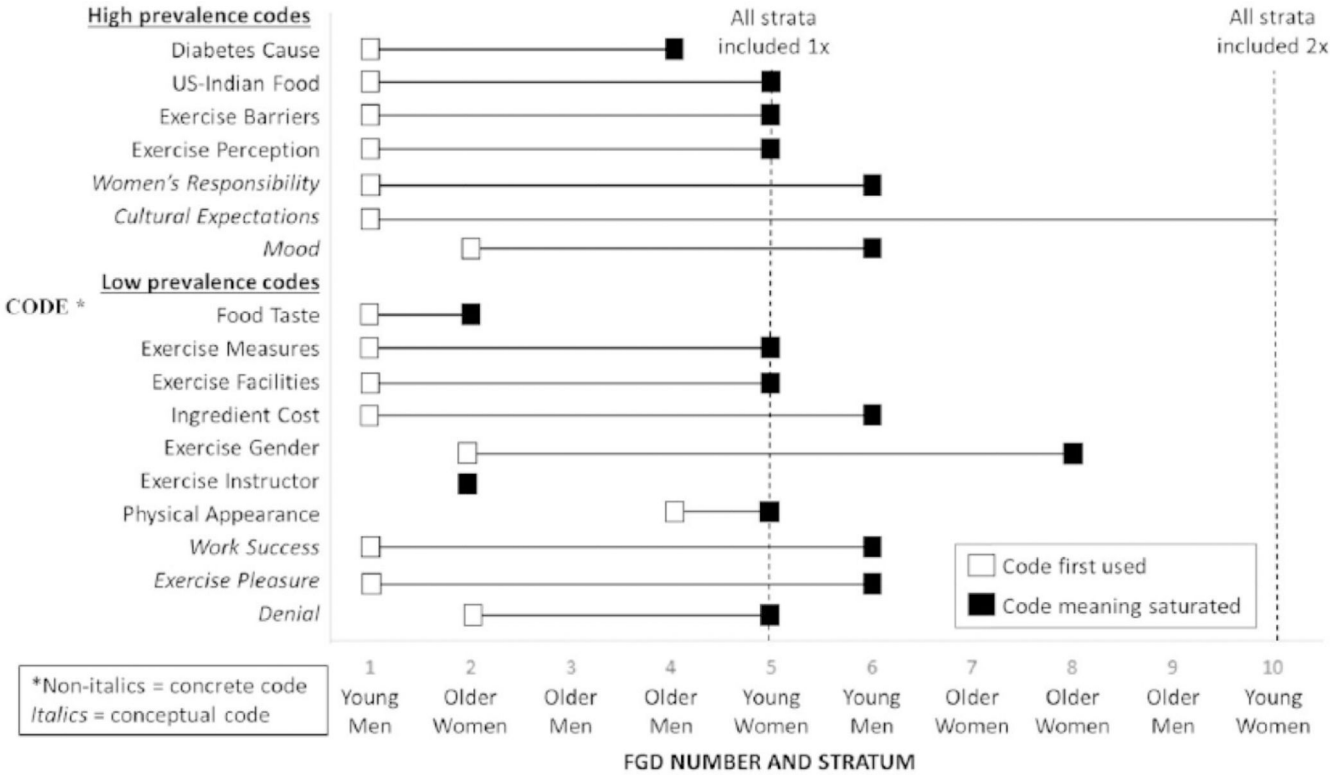

Figure 4.

Timing of first use of codes and their meaning saturation.

* concrete code.

Table 1.

Changes to codebook

	New Codes Created	Code Definitions Changed (Total)	Expanded Conceptually	Edited Inclusion/Exclusion Criteria	Added Examples
FGD 1	30	-	-	-	-
FGD 2	12	13	6	1	8
FGD 3	2	10	4	-	6
FGD 4	3	3	1	-	2
FGD 5	1	3	3	-	-
FGD 6	2	-	-	-	-
FGD 7–10	-	-	-	-	-
Total	50	31	14	1	16
Total (inductive codes)	29 (58%)	5 (16%)	4 (29%)	-	1 (94%)
Total (deductive codes)	21 (42%)	26 (84%)	10 (71%)	1 (100%)	15 (6%)

Table 2.

Examples of code dimensions identified across demographic strata of focus group discussions

Code	FGD 1 Strata: Young Men	FGD 2 Strata: Older Women	FGDs 3 & 4 Strata: Older Men	FGD 5 Strata: Young Women
Exercise Barriers (concrete code)				
Dimensions raised across strata	Lack of time for exercise. Lack of exercise facilities. Exercise conflicts with family time. Cost of exercise activities.	Same issues repeated	Same issues repeated	Same issues repeated
Dimensions raised in specific strata	Little interest in physical appearance. Education prioritized over physical activity.	Little awareness of health benefits of exercise vs. weight loss.	Socializing valued over exercise. Exercise routine is challenge. Weather limits outdoor exercise.	No family encouragement. Need accompaniment. Home exercise not effective.
Mood (conceptual code)				
Dimensions raised across strata.	Laziness to exercise	Same issue repeated	Same issue repeated	Same issue repeated
Dimensions raised in specific strata	Longing for family influences diet. Cravings for traditional foods. Satisfaction of food after long work hours.	Apathy for diet once children grown	Stress eating influences diet. Mental calm influences eating.	Eating habits difficult to change.

Qual Health Res. Author manuscript; available in PMC 2019 August 01.

Table 3.

Parameters influencing saturation and sample size for focus group discussions

Parameter of Saturation	How Parameter Influences Sample Size
Study Purpose	A study where researchers aim to *identify* core issues in data requires a smaller sample size to reach saturation (e.g. 4 focus groups); while a study where researchers aim to *understand* the issues requires a larger sample size and is dependent on other parameters.
Type of Codes	A study where researchers seek to capture explicit, concrete codes will require a smaller sample size than where researchers seek to identify more complex, conceptual, and nuanced issues.
Group Stratification	A study where focus groups are not stratified by any characteristics will require a smaller sample size to reach saturation (e.g., 3–6 groups) than one where focus groups are stratified by specific characteristics - whereby enough groups to include all strata at least once are needed to reach saturation.
Groups per Strata	A study using stratified focus groups requires two groups per strata to reach meaning saturation; however, there is limited additional benefit for data richness in conducting more than two groups per strata.
Type of Saturation	A study seeking *code saturation* requires a smaller sample size (e.g. 4 focus groups) than a study seeking *meaning saturation* (e.g. 5+ focus groups).
Degree of Saturation	A study where researchers seek to reach 80% saturation will require a smaller sample size (e.g. 2–3 focus groups) than where researchers seek to reach 90% saturation (e.g. 4–5 focus groups).