Universiteit Utrecht

Innovation Studies Utrecht (ISU)

Working Paper Series

(I Can’t Get No) Saturation: A Simulation and Guidelines for Minimum Sample Sizes in Qualitative Research

Frank van Rijnsoever

ISU Working Paper #15.05

(I Can't Get No) Saturation: A Simulation and Guidelines for Minimum Sample Sizes in Qualitative Research

Abstract

This paper explores *the sample size in qualitative research that is required to reach theoretical saturation.* I conceptualize a population as consisting of sub-populations that contain different types of information sources that hold a number of codes. Theoretical saturation is reached after all the codes in the population have been observed once in the sample. I delineate three different scenarios to sample information sources: "random chance," which is based on probability sampling, "minimal information," which yields at least one new code per sampling step, and "maximum information," which yields the largest number of new codes per sampling step.

Next, I use simulations to assess the minimum sample size for each scenario for systematically varying hypothetical populations. I show that theoretical saturation is more dependent on the mean probability of observing codes than on the number of codes in a population. Moreover, the minimal and maximal information scenarios are significantly more efficient than random chance, but yield fewer repetitions per code to validate the findings. I formulate seven guidelines for purposive sampling and recommend that researchers follow a minimum information scenario.

Keywords

qualitative research, purposive sampling, theoretical saturation, sample size, simulation

1. Introduction

Qualitative research is becoming an increasingly prominent way to conduct scientific research in business, management, and organization studies (Bryman & Bell, 2011). In the first decade of the twenty-first century, more qualitative research has been published in top American management journals than in the preceding 20 years (Bluhm, Harman, Lee, & Mitchell, 2011). Qualitative research is seen as crucial in the process of building new theories (Bluhm et al., 2011; Eisenhardt, 1989; Eisenhardt & Graebner, 2007) and it allows researchers to describe how organizational or industrial change processes unfold over time (Langley, Smallman, Tsoukas, & Van de Ven, 2013; Poole, Van de Ven, & Dooley, 2000). Moreover, it gives close-up and in-depth insights into various organizational phenomena (Birkinshaw, Brannen, & Tung, 2011; Gephart, 2004) perspectives and motivations for actions (Bryman & Bell, 2011; Gephart, 2004). However, despite the explicit attention of journal editors to what qualitative research is and how it should be conducted (Gephart, 2004; Suddaby, 2006; Walsh et al., 2015), it is not always transparent how particular research was actually conducted (Bluhm et al., 2011; Gioia, Corley, & Hamilton, 2013). A typical topic of debate is what the size of a sample should be for qualitative research to be credible and dependable (Patton, 1990; Suddaby, 2006)[1]

A general statement from grounded theory about sample size is that the data collection and analysis should continue until the point at which no new codes[2] or concepts emerge (see Bryman, 2013; Coyne, 1997; Glaser and Strauss, 1967). This does not only mean that no new stories emerge, but also that no new codes that signify new properties of uncovered patterns emerge (Charmaz, 2014). At this point, "theoretical saturation" is reached; all the relevant information that is needed to gain complete insights into a topic has been found (Bryman, 2013; Bryman & Bell, 2011; Suddaby, 2006).

Most qualitative researchers who aim for theoretical saturation do not rely on probability sampling. Rather, the sampling procedure is purposive (Coyne, 1997; Marshall, 1996). It aims "to select information-rich cases whose study will illuminate the questions under study" (Patton, 1990, p. 169). The researcher decides which cases to include in the sample based on prior information like theory or insights gained during the data collection.

However, the minimum size of a purposive sample needed to reach theoretical saturation is difficult to estimate (Baker & Edwards, 2012; Bowen, 2008; Francis et al., 2010; Guest, Bunce, & Johnson, 2006; O'Reilly & Parker, 2012; Sandelowski, 1995; Suddaby, 2006).

There are two reasons why the minimum size of a purposive sample deserves more attention. First, theoretical saturation seems to call for a "more is better" sampling approach, as this minimizes the chances of codes being missed. However, the coding process in qualitative research is laborious and time consuming. As such, especially researchers with scarce resources do not want to oversample too much. Some scholars give tentative indications of sample sizes that often lie between 20 and 25 and are usually below 50 (see Mason, 2010 for an overview), but the theoretical mechanism on which these estimates are based is unknown.

Second, most research argues that determining whether theoretical saturation has been reached remains at the discretion of the researcher, who uses her or his own judgment and experience (Sandelowski, 1995; Suddaby, 2006; Trotter II, 2012; Tuckett, 2004). Patton (1990) even states that "there are no rules for sample size in qualitative inquiry" (p. 184). As such, the guidelines for judging the sample size are often implicit. The reason for this is that most qualitative research is largely an interpretivistic endeavor (Blaikie, 2007) that requires flexible creative thinking, experience, and tacit knowledge (Suddaby, 2006). This complicates the assessment of the credibility and dependability of many qualitative studies, even though this is deemed important in the field of management (Gephart, 2004; Gioia et al., 2013; Pratt, 2008), but also in other fields (Bryman, 2013; Lincoln & Guba, 1986; Patton, 1990). Moreover, not all researchers have the required experience to assess intuitively whether theoretical saturation has been reached. For them, articulating the assessment criteria in a set of guidelines can be helpful.

In this paper I explore the sample size that is required to reach theoretical saturation in various scenarios and I use these insights to formulate guidelines about purposive sampling. Following a simulation approach, I assess experimentally the effects of different population parameters on the minimum sample size. I first generate a series of systematically varying hypothetical populations. For each population, I assess the minimum sample sizes required to reach theoretical saturation for three different sampling scenarios: “random chance,” which is based on probability sampling, “minimal information,” which yields at least one new code per sampling step, and “maximum information,” which yields the largest number of new codes per sampling step. The latter two are purposive sampling scenarios.

I demonstrate that theoretical saturation is more dependent on the mean probability of observing codes than on the number of codes in a population. Moreover, when the mean probability of observing codes is low, the minimal information and maximum information scenarios are much more efficient in reaching theoretical saturation than the random chance scenario. However, the purposive scenarios yield significantly fewer multiple observations per code that can be used to validate the findings.

This study adds to earlier studies that base their sample size estimates on empirical data (Bowen, 2008; Marshall, 1996). By simulating the factors that influence the minimum purposive sample size, I give these estimates a theoretical basis (Davis, Eisenhardt, & Bingham, 2007). Moreover, my simulations show that the earlier empirical estimates for theoretical saturation are reasonable under most purposive sampling conditions.

Based on these analyses, I offer a set of guidelines that researchers can use to estimate whether theoretical saturation has been reached. These guidelines help to make more informed choices for sampling and add to the transparency of the research, but are by no means intended as mechanistic rules that reduce the flexibility of the researcher (Suddaby, 2006; Walsh et al., 2015).

In section 2 I discuss the theoretical concepts about purposive sampling. Section 3 describes the methods, and the results of my simulations are presented in section 4. In section 5, I draw conclusions, offer recommendations, and discuss the limitations.

2. Theoretical Concepts

I base this section largely on the existing literature on purposive sampling. I also introduce some new ideas that are sometimes implied by the literature, but that were never conceptualized.

2.1. Populations, Information Sources, and Sampling Steps

A *population* is the “universe of units of analysis” from which a sample can be drawn (Lewis-Beck, Bryman, & Futing Liao, 2004, p. 834). However, in qualitative research, the unit of analysis does not have to be the same as the unit from which information is gathered. For example, in case study research, the units of analysis are the cases under examination, but the information comes from sources like interviews or documents. I refer to these sources as *“information sources.”* In the context of interviews, information sources are often referred to as informants (Malterud, 2001; Marshall, 1996), but they can be any source that informs the researcher: other examples are sites to collect observational data, existing documents, or archival data. I refer to the total set of information sources that are potentially relevant to answering the research question as the population.

From this population, one or multiple information sources are sampled as part of an iterative process that includes data collection, analysis, and interpretation. At each iteration, the researcher has the opportunity to adjust the sampling procedure and to select a new information source to be sampled. I assume in this paper that at each iteration only one source is sampled; this assumption has no further consequences for the remainder of the paper. Moreover, I use the term *sampling steps* rather than iterations, as this excludes analysis and interpretation. Finally, too ease interpretation, but in contrast to formal quantitative sampling terminology, I count as sampling steps only the information sources that participated in the research, thus excluding non-respondents.

2.2. Sub-populations

A population of information sources is usually not homogeneous. Multiple *sub-populations* can often be distinguished, for example the difference between interviewees, documents, or focus groups. This is important as the researcher can choose different sampling procedures and data collection methods for each sub-population. The exact delineation of sub-populations depends on the judgment of the researcher, for which arguments need to be provided. However, I impose a number of restrictions on the delineation of sub-populations.

- First, information sources should be interchangeable at the sub-population level. Within a sub-population, no single information source may be critical for reaching theoretical saturation. Hence, no single information source in a sub-population can contain information that is not found in other information sources in that sub-population. The reason for this criterion is that if a particular information source is critical for theoretical saturation, it should by definition be included in the research. Observing critical information is not guaranteed if the inclusion is dependent on a particular sampling strategy. A critical information source should then be treated as a separate sub-population of size one.
- Second, if cases or groups are compared, it is important to treat these as sub-populations. For example, distinguishing between sub-populations is a condition for data triangulation, because the researcher effectively compares the results from one sub-population (for example interviews with managers) with the results from another (for example annual reports). Furthermore, comparative case studies (Eisenhardt & Graebner, 2007; Yin, 2009) involve the comparison of sub-populations.

- Third, if there are differences in the type of information source, sampling strategy, data collection, or methods of analysis, then there are sub-populations. The reason for this criterion is that different methods are needed. These different methods need to be accounted for (Lincoln & Guba, 1986) as they can explain differences in outcomes.

The concept of sub-populations implies that theoretical saturation can be reached at the level of the overall population or at the level of the sub-population. Reaching theoretical saturation in all the sub-populations is not a condition for reaching theoretical saturation at the level of the population, since sub-populations can have an overlap in information. Where to strive for theoretical saturation is at the discretion of the researcher. However, it is necessary to reach theoretical saturation in each sub-population in comparative research, as this is the only way to make a valid comparison.

2.3. Codes, Mean Observations, and Repetitiveness

In most cases of inductive qualitative research, information is extracted from information sources, interpreted and translated into *codes*. I refer to codes here in the context of grounded theory, which means that they can be seen as “tags” or “labels” on unique pieces of information (Bryman, 2013). Codes can represent any sort of information and may be related to each other (for example phenomena and explanations). The only conditions that I impose are that each code represents only one piece of information and that two different codes are not allowed to represent the same information (e.g. synonyms are removed). Thereby, codes can be interpreted as unique “bits” of information.

The population contains all the codes that can be potentially observed. At the start of a study, the codes in the population are unobserved and the exact number of codes in the population is unknown. Consulting information sources sampled from the population allows codes to become observed. Theoretical saturation is reached when each code in the population has been observed at least once.

The number of sampling steps required to reach theoretical saturation depends on two population characteristics. First, the larger the *number of codes* distinguished in the population, the more sampling steps are required to observe them all. Second, the more often a code is present in the population, the larger are the chances that it will become observed. As theoretical saturation takes place at the population level, the distribution of codes in the population is important. For example, interviews can vary in length or some documents can contain more relevant information than others. In general, one would expect that the higher the *mean probability of observing codes* in a population is, the fewer sampling steps are required to reach theoretical saturation. This means that as the probability of observing codes increases in the population, the smaller the sample needs to be.

Purposive sampling allows the researcher to make an informed estimation about the probability of observing a given code at each sampling step, using theoretical prior information or insights gained during the analysis data so far[1]. However, when the number of codes is large, it is easier simply to estimate the mean probability of observing all the codes in the population and the variance. To make such estimations, it is important to consider what the probability of observing codes actually represents. The probability of a code being present at least depends on (1) the likelihood of an information source actually containing the code, (2) the willingness and ability of the source (or its authors) to let the code be uncovered, and (3) the ability of the researcher to observe the code. These probability estimations are based on the characteristics of the information source and the researcher. The probability of observing codes decreases when the information source (for example an interviewee) has strategic reasons not to share information. This happens for example in the analysis of corporate social responsibility interviews with firms (Shnayder, Van Rijnsoever, & Hekkert, in press). In addition, if the researcher has less experience with the technique used to uncover codes from a source or with correctly interpreting information during the data analysis, the probability of observing codes decreases.

Grounded theory considers codes that are observed more than once as redundant, since they do not add new information to the data (Lincoln & Guba, 1986; Patton, 1990; Trotter II, 2012). I refer to codes that are observed more than once more neutrally as "*repetitive codes*." Repetitive codes are important for a methodological purpose: they can help guard against misinformation. That is, information sources may have given false codes, for reasons of social desirability, strategy, or accidental errors. Moreover, the researcher may have misinterpreted information from a source. In all these cases, a code will receive the value 1, when it should have been 0.

To guard against misinformation and to enhance the credibility of the research, it can be advisable to aim for a sample in which each code is observed multiple times (this also follows from the logic behind triangulation). One could argue that if a code, after a substantial number of sampling steps, is still observed only once while almost all other codes have a higher incidence, a critical examination of the code is warranted. In many cases, the researcher may already be suspicious of such a code during the analysis. A frequency of one does not mean that the code is wrong by definition; it is possible that the code is just rare or that the low frequency is just a coincidence. However, it is relatively easy to make an argumentative judgment about the plausibility of rare codes (for example based on theory).

2.4. Sampling Strategies, Sampling Scenarios, and Efficiency

[1] This conceptualizing of purposive sampling is also consistent with the notion of theoretical sampling. Both terms are often used interchangeably. Theoretical sampling can be seen as a special case of purposive sampling (Coyne, 1997).

A *sampling strategy* describes how the researcher selects the information sources. Examples are “maximum variation sampling,” “typical case sampling,” and “snowball sampling.” The most elaborate inventory comes from Patton (1990), who identifies 15 purposive sampling strategies for qualitative research (also see Bryman, 2013). These strategies are based strongly on research practices, but the underlying theoretical criteria for distinguishing between the strategies are left implicit[3].

I use the concepts described above to formulate three generic *sampling scenarios*. I refer to sampling scenarios to avoid confusion with the aforementioned strategies. The term scenarios term signifies that they are based on theoretical notions, instead of empirical data or observed practices. The three sampling scenarios are based on the number of newly observed codes that a sampled information source adds. This criterion is motivated by the premise of purposive sampling: based on the expected information, the researcher makes an informed decision about the next information source to be sampled at each sampling step. The researcher can thus reasonably foresee whether, and perhaps how many, new codes will be observed at the next sampling step. The fewer sampling steps that a scenario requires to reach theoretical saturation, the more *efficient* it is.

The three scenarios that I identify are “random chance,” “minimal information,” and “maximal information.”

- *Random chance* assumes that the researcher does not use prior information during each sampling step. The researcher randomly samples an information source from the population and adds it to the sample. This scenario is solely based on probability and is considered to be inappropriate for most qualitative studies (Coyne, 1997; Marshall, 1996). However, there are good reasons to include this scenario. First, there are conditions under which random chance is an appropriate scenario for sampling. One of these is when no information is gained about the population during the sampling steps, such as when documents or websites are analyzed. Second, random chance can be seen as a worst-case scenario. If a researcher is uncertain about how a sampling process actually worked, it is always possible to explore whether theoretical saturation would have been reached under the conservative conditions of random chance. Third, random chance is the only scenario for which the number of sampling steps can be calculated mathematically. Finally, the random chance scenario can serve as an objective benchmark to which the number of sampling steps in the other scenarios can be compared.
- *Minimal information* is a purposive scenario that works in the same way as random chance, but adds as extra condition that at least one new code must be observed at each sampling step. This is equivalent to a situation in which the researcher actively seeks information sources that reveal new codes, for example by making enquiries about the source beforehand. It is not uncommon for a researcher to discuss topics with a potential interviewee prior to the actual interview to assess whether the interview will be worthwhile. The minimal information scenario captures these kinds

of enquiries. Similarly, researchers may be referred to a next source that adds new codes as part of a snowball strategy. Overall, the criterion of observing at least one new code per sampling step seems to be relatively easy to achieve as long as the researcher has some information about the population at each step. This makes the scenario broadly applicable and more efficient than random chance.

- *Maximal information* is a purposive scenario that assumes that the researcher has almost full knowledge of the codes that exist in the population and the information sources in which they are present. At each sampling step, an information source is added to the sample that leads to the largest possible increase in observed codes. Thereby, this scenario is most in line with the theoretical aim of purposive sampling. It is likely to be the most efficient scenario, but it also makes large assumptions regarding the prior knowledge of the researcher about the population. An example of when this scenario might be realistic occurs when the researcher tries to replicate results or when the researcher is extremely familiar with the field and the specific setting that he or she is investigating. However, this scenario is unrealistic when the populations are large and not well defined beforehand. An example of this is interviewing entrepreneurs in a given region. There are often many potential respondents and these are often not properly registered, making it difficult to gain full information about the population.

I use these scenarios to simulate the number of sampling steps required to be likely to reach theoretical saturation.

3. Methods

I first provide formal notation of the concepts described above. These are used to simulate the three scenarios with regard to the number of sampling steps necessary to achieve theoretical saturation and the number of repetitive codes.

Simulations allow me to assess the effects of the three scenarios on a series of hypothetical populations that vary systematically regarding (1) *the number of codes in the population* and (2) *the mean probability of observing codes*. The controlled setting allows me to assess the relative influence of each of these factors on the reaching of theoretical saturation. In an empirical setting, this would not be possible, because the researcher can generally not control the characteristics of the population under study, because the number of populations that can be studied is limited and because it is never entirely certain whether theoretical saturation has been reached (Blaikie, 2007).

3.1. Definitions

The principles discussed below are applicable to the population level if there are no sub-populations. If there are sub-populations, everything can be applied to that level. I denote the population by *J* and a given information source by *i.* The number of sampling steps is given by *n*, and the number of sampling steps required for theoretical saturation is given by n_s. Moreover, I denote a sub-population by *j* and the number of sub-populations by *m*. When working at the sub-population level, everything needs to be subscripted with *j*. However, for convenience of notation, I work here at the population level.

The number of codes in the population is denoted by k. The vector of codes in the population is given by *C,* which has length k:

- $C = (C_1, C_2, \dots, C_k)$

In the case that the vector *C* represents all the codes in the population, it is subscripted with *J*. By definition, the vector C_J has length *k* and all the values are 1 (all the codes are present, but still unobserved). In addition, each information source *i* in the population consists of a vector of codes denoted by c_i that is also of length *k*. It indicates whether a code is present (1) or absent (0) in a source:

- $c_i = (c_{i1}, c_{i2}, \dots, c_{ik})$

The vector C_{fn} is also of length *k* and contains how often (e.g. the frequency) each code has been observed after *n* sampling steps. The researcher can use the vector C_{fn} to assess the credibility of the research by looking at rare codes. It is easily obtained by calculating the sums of all vectors c_i.

- $C_{fn} = \sum_{i=1}^{n} c_i$

The mean of C_{fn} is denoted by $\overline{C_{fn}}$; at n_s it is denoted by $\overline{C_{fn_s}}$. Further, I denote a vector C_{On}, also of length *k*, that expresses which codes have been observed after *n* sampling steps. It is obtained by substituting all the values of C_{fn}>0 with 1. Theoretical saturation is reached if:

- $C_J = C_{On}$

The presence or absence of an example code c_1 in an information source *i* can be seen as a random Bernoulli trial. The probability that c_1 is present (1) is given by Φ_{c1} and the probability that it is absent (0) is given by 1-Φ_{c1}. The vector Φ_c of length k contains the probabilities of each code being present:

- $\Phi_{\mathrm{c}} = (\Phi_{\mathrm{c1}}, \Phi_{\mathrm{c2}} \dots, \Phi_{\mathrm{ck}})$

The vector Φ_c can be described by a beta distribution, which has two parameters that can only take a value larger than 0: α and β. The *expected mean* of the distribution is given by:

- $E[\Phi_c] = \frac{\alpha}{\alpha+\beta}$

The variance of the beta distribution lies between 0 (all the values are the same) and 0.25 (exactly half of the values are 0 and half of the values are 1). The expected variance is given by:

- $Var[\Phi_c] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

The expected mean of the distribution is a theoretical construct that I use in my simulations. However, qualitative researchers usually deal with the observed distribution of probabilities, which I will denote as $\overline{\Phi_c}$. It is obtained by calculating the mean of vector Φ_c. $E[\Phi_c]$ and $\overline{\Phi_c}$ are expected to have a high – but not perfect – correlation, because observed (empirical) distributions always contain deviations that result from random chance.

To answer the research question, I need to know the value of *n* for *F4* to be true with a certain probability, which I denote by p_n. It can be seen as the chance that theoretical saturation has been reached after *n* sampling steps. The larger the value of p_n, the larger *n* needs to be, given *k* and $\overline{\Phi_c}$. Consistent with the conventions in quantitative research, I choose to set p_n at 0.95. This means that I look for a value of *n* for which all the codes are observed in 95% of the times that I draw a sample. The value of 95% is arbitrary, but it gives a degree of certainty that is equal to the common significance threshold in the social sciences.

In the case of random chance, the probability p_{nc1} that example code c_1 is actually observed at least once after *n* sampling steps is equal to 1 minus the probability that the code has not been observed after *n* sampling steps:

- $p_{nc1} = 1 - (1 - \Phi_{c1})^n$

In the case of *k* codes, the probability p_n of observing all the codes is the product of each probability from *F8*:

- $p_n = \prod_{c=1}^{k}(1 - (1 - \Phi_{ck})^n)$

A special condition occurs when all the values of Φ_c are equal, which implies a variance of zero. The larger the values of α and/or β, the more this condition is approximated. The value of Φ_c can then be denoted as a single number Φ_k. Though not very realistic, this special condition is mathematically convenient, since it is the only case in which we can calculate *n_s* directly. This allows me to explore the relationship between n_s, k, and Φ analytically under the conditions of random chance. First, I can simplify *F9* to:

- $p_n = (1 - (1 - \Phi_k)^n)^k$

This can be rewritten in terms of *n*. Moreover, I want to obtain the value of n_s, which is a specific value of *n*:

- $n_s = \frac{\ln(1 - \sqrt[k]{p_n})}{\ln(1 - \Phi_k)}$

This formula gives the number of sampling steps needed to reach theoretical saturation in 100*p_n% of the cases, based on random chance, as a function of *k* and Φ_k. If p_n is set at 0.95, then the number of required sampling steps is formally denoted as $n_{s(0.95)}$. However, since p_n is fixed throughout this paper at 0.95, the addition of the number has little added value. For convenience of notation, I continue to refer to n_s.

If the values of Φ_c are not equal, it is not possible to write *F9* in terms of n_s, which means that the minimum sample size cannot be calculated mathematically. In that case, simulation techniques can be used to approximate the answer (Law & Kelton, 1991).

To enhance the credibility of the research, it is also possible to aim deliberately for a minimum number of observations of each code. This minimum is denoted by *ν* (Greek letter nu). A value of *ν*=2 means that each code is observed at least twice and thus that each code is repeated once. The probability of observing *k* codes, with probability Φ_{ck}, for a *ν* number of times after *n* sampling steps is given by:

- $p_n = (\prod_{k=1} 1 - (1 - \Phi_{ck})^n)^\nu$

In the specific condition that all the values of Φ_c are equal, this simplifies to:

- $p_n = (1 - (1 - \Phi_k)^n)^k)^\nu$

Rearranging and replacing *n* with n_s gives:

- $n_s = \frac{\ln(1 - \sqrt[k\nu]{p_n})}{\ln(1 - \Phi_k)}$

F14 provides, based on random chance, the minimum number of sampling steps to observe each code *ν* times, conditional on the other variables. Again, in the case that the values of Φ_c are not equal, it is not possible to write *F12* in terms of n_s.

Finally, if these principles are applied to *m* sub-populations, the probability of reaching theoretical saturation at the population level is calculated by:

- $p_n = \prod_{m=1} p_{n_j}$

In the case that all the probabilities of reaching theoretical saturation are set as equal for all the sub-populations, *F15* simplifies to:

- $p_n = (p_{n_j})^m$

If there are three sub-populations and p_{n_j} in each of these is set at 0.95, then the probability of reaching theoretical saturation in the overall population p_n is 0.857. However, unless the purpose is to compare sub-populations, this number should be considered as a minimum probability. Since sub-populations can have an information overlap, codes that have not yet been observed in one sub-population may have already been observed in another.

3.2. Simulation of Scenarios

As a benchmark, I first calculate *n* under the assumption that the $Var[\Phi_c]=0$, using *F11*. Next, using the R-program (R Development Core Team, 2015), I simulate 1100 hypothetical populations of information sources. The populations vary systematically by the number of codes (*k*) from 1 to 101 with increments of 10, for all discrete combinations of *α* and *β* that lie between 1 and 10. Each information source in a population is represented by a vector c_i. The size of each population is 5000 information sources, which means that each population is actually a 5000 by *k* matrix, with only values of 0 and 1. I choose a population size of 5000 to prevent this number from influencing the results.[4] For smaller populations, theoretical saturation is likely to be reached earlier, but I take a conservative approach. Further, in line with my earlier argument about interchangeability of information sources, I impose a condition whereby each of the *k* codes is actually present in at least two information sources in the population.

By setting the *α* and *β* parameters to >1, I obtain only unimodal distributions and exclude all U-shaped or J-shaped distributions. These shapes imply that within one distribution, the probability of observing codes that are rare or common is larger than the probabilities of any other code being observed. U-shaped or J-shaped distributions are not impossible but less plausible than a distribution with only increasing or decreasing probabilities. The expected means of the simulated populations lie between 1/11 and 10/11 and the maximum variance is 1/12 (at *α*=*β*=1).

For each population, I simulate the number of sampling steps necessary to reach theoretical saturation under the three scenarios from section 2.5. Figure 1 gives a schematic overview of how the algorithms for each scenario operate.

---Insert Figure 1 about here---

All three scenarios operate in a similar manner. After generating a population, an information source is selected, but the selection process differs per scenario:

- *Random chance* selects information sources based on probability.
- *Minimal information* works in the same way as random chance, but adds an extra condition that *i* must contain at least one code that is not already in C_{On} (the vector indicating which codes are observed), otherwise *i* is discarded.
- *Maximal information* first identifies a set of information sources J_{max} that add the largest number of new codes to C_{On}. From this set, which often consists of a small number of information sources, it randomly selects an information source.

If the source has not been selected before, it is added to the sample. After each sampling step, the vector C_{On} is updated with the new codes that are found in c_i. Next, C_{On} is compared with C_J; if the two are equal, then theoretical saturation is reached in the population. Consequently, the process stops and n_s is reported. Otherwise, the next sampling step takes place and a new information source is selected from the population.

I apply each of the three sampling scenarios to each population 500 times. This produces a distribution for each scenario with values of n_s. From this distribution, I obtain n_s. Moreover, for each population, I obtain the average value over 500 simulations of $\overline{C_{fn_s}}$ (the vector containing the frequencies of observed codes) as an indicator of repetitive codes. For convenience of notation, *F* represents a vector of length 500 containing the values of $\overline{C_{fn_s}}$. The average of this vector is given by $\bar{F}$: it is equal to the more elaborate notation $\overline{\overline{C_{fn_s}}}$ that I shall not use further.

4. Results

Figure 2 plots the 95th percentile of n_s (the number of sampling steps required to reach theoretical saturation) against $\overline{\Phi_c}$ (the mean probability of observing codes) for the values of *k* (the number of codes) between 11 and 101. Note that the y-axis is logarithmic. I do not show *k*=1, because for purposive scenarios it has a constant value of 1. The solid black line indicates the calculated random chance's value of *n* based on *F11*. The blue dots represent random chance, the green diamonds represent minimal information, and the red triangles represent maximal information. The random chance scenario generally follows the calculated line from *F11*, but requires more sampling steps than *F11*. This is due to the fact that a variance larger than 0 in Φ_c means that some codes have a probability of being observed that is smaller than $\overline{\Phi_c}$. Making the rare codes observed requires more sampling steps. This result demonstrates that random chance algorithm worked well.

---Insert Figure 2 about here---

Figure 2 shows that in the random chance scenario, the low values of $\overline{\Phi_c}$ lead to values of n_s that are larger than 4000, regardless of *k*. As $\overline{\Phi_c}$ increases, n_s declines rapidly with a decreasing trend to below 10 for all the values of *k*. This implies that $\overline{\Phi_c}$ is more important than *k* for reaching theoretical saturation. The figure also shows that both purposive scenarios are more efficient than random chance. For smaller values of $\overline{\Phi_c}$, the differences are the largest. With random chance for *k*=101 and $\overline{\Phi_c}$<0.1, it generally requires more than 1000 sampling steps to reach theoretical saturation in 95% of the cases. Under the same conditions, this number is reduced to about 46 information sources in the minimal information scenario and to about 20 in the maximal information scenario. As $\overline{\Phi_c}$ becomes larger, the random chance and minimal information scenarios become about equally efficient, while the maximal information scenario remains more efficient. Notable is that the numbers of both purposive scenarios fall within the range of common indications of sample size from the literature (below 50). This result confirms that this indication is not far from accurate. Finally, in the maximal information scenario, n_s has little variance for values of $\overline{\Phi_c}$. This is because *n* can only be a discrete integer and because the high level of efficiency leads to small values of *n*, which does not give much room for variation.

Figure 3 plots $\bar{F}$ (the mean number of observations per code) upon reaching theoretical saturation for each scenario against $\overline{\Phi_c}$ for the values of *k* between 11 and 101. Again, the blue dots represent random chance, the green diamonds represent minimal information, and the red triangles represent maximal information.

---Insert Figure 3 about here---

In line with the result above, $\overline{\Phi_c}$ has a greater influence than *k* on $\bar{F}$ in the random chance scenario. This is not surprising as n_s and $\bar{F}$ have a correlation of 0.902 in the random chance scenario. Second, the random chance scenario gives the largest number of repetitive codes (over 400) at low values of $\overline{\Phi_c}$. This is explained by the fact that it has the most sampling steps on average. However, for higher values of $\overline{\Phi_c}$, random chance yields about the same number of codes as minimal information, which is between 3 and 5. In addition, the relationship between $\overline{\Phi_c}$ and $\bar{F}$ is slightly curvilinear in the minimal information scenario. This finding is explained by the fact that when $\overline{\Phi_c}$ is low, vector c_i often only contains one code. Since each step must lead to a new code to become observed, there are few repetitive codes. When $\overline{\Phi_c}$ becomes larger, vector c_i is more likely to contain multiple codes, which increases the chances of observing codes multiple times. However, as the number of observed codes per information source increases, the minimal information scenario becomes more efficient with regard to reaching theoretical saturation. At high values of $\overline{\Phi_c}$, this efficiency effect becomes stronger than the effect of having multiple codes per source. Finally, the maximal information scenario only yields between 1 and 3 observations per code and it is characterized bya series of positive lines at higher values of $\overline{\Phi_c}$. These are the result of the discrete character of *n*. As $\overline{\Phi_c}$ increases, *n* remains stable until its next lower integer, but it does increase gradually. The low number of codes and these patterns makes the use of repetitive codes for maximal information very limited.

Overall, the results show that there is a clear trade-off between the efficiency of the scenario and the number of repetitive codes. To increase the credibility of the research, it is possible to aim for a minimum number of observations of each code (*v*). For reasons of space, I do not simulate the various scenarios for different values of *v*, but *F14* shows that it is relatively easy to aim for observing codes multiple times. On average, to obtain a repetition of one (*v*=2) based on the calculated random chance, 2.3 extra sampling steps are required, which is an increase of about 10%. For *v*=3, 3.66 extra steps are required (about 17%), and for *v*=4, 4.66 extra steps are required (about 21%). There is a large variation in absolute numbers; for example, for *k*=101 and $\overline{\Phi_c}$ =0.09, the number of extra steps is 68.95. However, compared with the *n* required to reach theoretical saturation under these conditions (see *F11*), this is only a 9% increase. Setting *v*>1 thus requires only a limited number of extra sampling steps. The number of extra steps required is even smaller for both purposive scenarios as these are more efficient.

5. Conclusions

The results for the purposive scenarios produced the same range of minimum sample sizes (below 50 information sources) as tentatively indicated in the literature. The simulations also uncovered mechanisms that give key insights into the estimation of the minimum size of a qualitative sample. The mean probability of observing codes is more important than the number of codes in the population for reaching theoretical saturation. Furthermore, when the probability of observing codes is low, the purposive scenarios are much more efficient than the random chance scenario. When this probability is high, the differences between scenarios are small. Finally, the more efficient a scenario is, the lower the mean number of observations per code, but only a few sampling steps are required to increase the minimum number of observations of all the codes.

5.1. Limitations and Further Research

This paper has two potential limitations that deserve discussion. First, critics could claim that the scenarios are mechanistic and do not represent real-world sampling procedures. I used ideal typical scenarios that capture the full range of possible empirical sampling procedures. Researchers who view their research through the lens of these scenarios are likely to observe that their sampling procedure shares characteristics with at least one of the three scenarios or that their sampling procedure is a mixture of two scenarios. Future researchers can also simulate other scenarios that they conceive and even include different sampling strategies in their simulations, like snowball sampling or sampling for maximal variation (Bryman, 2013; Patton, 1990).

Second, I simulated a broad range of scenarios for the purpose of this paper, but other simulations are also possible. For example, I simulated only one population per combination of α, β, and the number of codes *(k)*. This lack of variation could cast doubt on the robustness of my results. However, it should be noted that the 1100 populations did have a large variation, as the number of codes was not important for the minimum sample size and because the different combinations of α and β often led to the same mean probability of observing codes (e.g. the variance was not important). By setting α and β to larger than one, I also only excluded all J- and U-shaped distributions of code probabilities since I consider them unrealistic in an empirical setting. I also did not vary the population sizes. Instead, I chose a large number that produced conservative estimates of the minimal sample size. It would be empirically interesting to vary the sample sizes in the simulations. For computational reasons and to reduce the complexity of this paper, I left this challenge for future researchers. Finally, I did not simulate different minimum observations per code, as the formula based on random chance gave sufficient insights into this issue.

5.2. Guidelines for Purposive Sampling

Based on these insights, I formulate a set of guidelines for sampling in qualitative research. These are not intended as formal mechanistic rules, but rather as an aid to making informed choices about the sampling and how it is reported.

The guidelines for sampling in qualitative research consist of the following seven steps:

1. *Identify a population of information sources and sub-populations.* This does not need to be a formal sampling frame, but the researcher does need to sketch the kind of information sources that exist in the population and whether there are sub-populations. If there are sub-populations, the researcher needs to argue:
    a. The basis for distinguishing sub-populations.
    b. Whether the sources are interchangeable in a sub-population.
    c. Whether the sub-populations serve a comparative purpose or are used for other means.
    d. The process of data collection, sampling, and analysis per sub-population.
    e. Other criteria that are deemed important by the researcher.

    The more detailed the researcher's description of the population and sub-populations, the better. This is especially true when the researcher aims to use a maximal information scenario. However, as the researcher should keep an eye open for new developments, the delineation of the population and sub-populations can be updated at each sampling step.
2. *Estimate the number of codes per sub-population.* This estimation is based on:

    a. The complexity and scope of the research question.
    b. The existing theory and information available about the sub-population.
    c. Other possible factors that are deemed to be of influence.

   Because the influence of the number of codes on theoretical saturation is small, it is more important to give an order of magnitude than an exact number. In most instances, assuming around 100 codes is safe, as this is a common number for observed codes and the minimum number of sampling steps is relatively stable in this order of magnitude. This estimation can be adapted after each sampling step.

3. *Estimate the mean probability of a code being observed.* The researcher does not need to know what a reasonable probability is at the start of the research, but it is likely that after consulting a number of information sources, the researcher will have enough information to make the assessment. The judgment at least depends on:
    a. The likelihood of an information source actually containing codes (is required information rare in the population and what are the chances of non-response?).
    b. The willingness and ability of the source (or its authors) to let the code be uncovered (are there strategic interests?).
    c. The probability that the researcher is able to observe the code (based on the researcher's prior research experience and familiarity with the topic).
    d. Other criteria that are deemed important by the researcher.
4. *Set a degree of certainty to reach theoretical saturation.* In line with statistical conventions, I used 0.95, but researchers should feel free to deviate.
5. *Assess which scenario is most applicable to each sub-population*.
    a. Random chance is only appropriate if after a substantial number of sampling steps, the researcher still has little or no idea about the characteristics of the sub-population and where codes can be found. In that case, random chance serves as a fallback scenario. If theoretical saturation is reached under random chance, then it is also reached in the other two scenarios. With conservative estimates of the mean probability of observing codes, the minimum sample size is over 4000 information sources, while for higher means, the minimum sample size rapidly drops to below 100 at probabilities of around 0.3 and below 50 at probabilities of 0.4.
    b. Choosing a minimal information scenario requires some argumentation. Most important is that the researcher makes it plausible that a new code will be observed at each sampling step. This is something that the researcher will experience as the research progresses. Usually there is little need to aim deliberately for multiple observations per code, because the scenario delivers sufficient repetitive codes. Only under very high or very low estimates of the mean probability of observing codes might the researcher consider aiming for at least two observations per code. Under low estimates of the mean probability of observing codes, the minimum sample size for minimal

information is around 50, while for higher means the minimum sample size is below 25.

c. The researcher can only choose maximum information when there is already a full overview of all the information sources in the (sub-)population and how information-rich these sources are (e.g. how many codes they contain). However, as maximum information makes very strong assumptions, the choice needs proper argumentation that is at least based on the criteria discussed in steps 2 and 3. The benefit of the maximum information scenario is that even under low estimates of the mean probability of observing codes, the minimum sample size is only 20 information sources. For higher means, the minimum sample size drops below 10. However, unless there is already strong theory present, it is advisable to aim for multiple observations of each code to guard against misinformation.

It is unlikely that a scenario will be followed exactly; rather, the researcher will notice that the sampling procedure falls somewhere in between the scenarios. As such, the researcher can argue which scenario the sampling procedure resembles most. The researcher can use the results from the simulations above to assess whether theoretical saturation is likely to have been reached.

6. *Choose a fitting sampling strategy*. The researcher should take into account that the sampling strategy needs to lead to a sufficiently broad reach across information sources in the population to be able to cover all the codes.
7. *Account for these steps when reporting the research.* State why a scenario, with its associated minimum sample size is appropriate. If applicable, the reasons for deviating from these steps are reported. The researcher also reports the number of times each code was observed and whether there were reasons to suspect that some codes were not credible. Finally, the researcher reports the probability of having reached theoretical saturation at the population level (see *F15* and *F16*).

Following these recommendations, does not mean that overall quality of the research is good. They can only help to improve the sampling, which is but one aspect of the entire process. In addition, in many instances, codes are not yet fixed at the start of the research. Rather, they become more known as the research progresses. The approach proposed here takes this into account by allowing researchers to reevaluate their assessment during each sampling step.

Keeping the analyses in mind, I recommend that researchers should generally opt for a minimal information strategy, as it makes reasonable assumptions, it is efficient, and it yields sufficient codes. Whether saturation has been reached remains in the argumentative judgment of the researcher. These guidelines can aid the researcher in making this judgment and the readers in assessing it. Overall, the results and the guidelines offered in this paper can improve the quality and transparency of purposive sampling procedures. Therefore, I encourage fellow researchers to consider using these ideas and guidelines and to improve upon them where they see fit.

[1] In this paper, I refer to qualitative research in an inductive context that is inspired by the ideas about grounded theory by Glaser and Strauss (1967). This approach has strongly influenced other forms of inductive qualitative research (Bryman, 2013). I recognize that there are more deductive-oriented forms of qualitative research.

[2] To prevent confusion, I use the term 'code' in this article to refer to information uncovered in qualitative research. The term 'concept' refers to the concepts in the theoretical framework.

[3] For example, a criterion that can explain the difference between "maximum variation sampling," "typical case sampling," and "extreme case sampling" is the focus of the research question. "Snowball sampling" and "opportunistic sampling" differ in the way in which they obtain information about the next information source that is to be sampled. "Confirming or disconfirming sampling" and "including politically sensitive cases" as strategies are motivated by a delineation of the population. Overall, Patton (1990) acknowledges that purposive sampling in qualitative research can be a mixture of the strategies identified and that some of these strategies overlap. These strategies also make implicit assumptions regarding the prior knowledge of the researcher about the population. For example, "extreme case sampling" implicitly assumes that the researcher has knowledge about the full population; otherwise, he or she would be not be able to identify the extreme cases. "Snowball sampling" assumes that the researcher does not have full knowledge of the population, as relevant leads are only identified at each sampling step.

[4] Technically, for the random chance and minimum information scenarios, there is no need to specify a population size a priori. One can just generate a new information source at each step. However, maximal information does require an a priori specified population, as it assumes that the researcher has full knowledge of the population. To keep the results comparable across scenarios, I also specify a population size for random chance and minimum information.

References

Baker, S. E., & Edwards, R. (2012). *How many qualitative interviews is enough*. *National Centre for Research Methods Review Paper*. Southhampton: National Centre for Research Methods.

Birkinshaw, J., Brannen, M. Y., & Tung, R. L. (2011). From a distance and generalizable to up close and grounded: Reclaiming a place for qualitative methods in international business research. *Journal of International Business Studies*, *42*(5), 573–581.

Blaikie, N. (2007). *Approaches to Social Enquiry*. Cambridge: polity.

Bluhm, D. J., Harman, W., Lee, T. W., & Mitchell, T. R. (2011). Qualitative research in management: a decade of progress. *Journal of Management Studies*, *48*(8), 1866–1891.

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

Bryman, A. (2013). *Social Research Methods* (4rd ed.). Oxford: Oxford University Press.

Bryman, A., & Bell, E. (2011). *Business Research Methods 3e*. Oxford: Oxford university press.

Charmaz, K. (2014). *Constructing grounded theory*. London, UK: Sage.

Coyne, I. T. (1997). Sampling in qualitative research. Purposeful and theoretical sampling; merging or clear boundaries? *Journal of Advanced Nursing*, *26*(3), 623–630.

Davis, J., Eisenhardt, K., & Bingham, C. (2007). Developing Theory Through Simulation Methods. *The Academy of Management Review*, *32*(2), 480–499.

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *The Academy of Management Review*, *14*(4), 532–550.

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory building from cases: opportunities and challenges. *Academy of Management Journal*, *50*(1), 25–32.

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 and Health*, *25*(10), 1229–1245.

Gephart, R. P. (2004). Qualitative research and the Academy of Management Journal. *Academy of Management Journal*, *47*(4), 454–462.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research notes on the gioia methodology. *Organizational Research Methods*, *16*(1), 15–31.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.

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.

Langley, A., Smallman, C., Tsoukas, H., & Van de Ven, A. H. (2013). Process studies of change in organization and management: unveiling temporality, activity, and flow. *Academy of Management Journal*, *56*(1), 1–13.

Law, A. M., & Kelton, W. D. (1991). *Developing Theory Through Simulation Methods* (Vol. 2). McGraw-Hill New York.

Lewis-Beck, M. S., Bryman, A., & Futing Liao, T. (2004). *The Sage Encyclopydia of Social Science Research Methods* (Vol. 1–3). London: Sage.

Lincoln, Y. S., & Guba, E. G. (1986). But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Directions for Program Evaluation*, *1986*(30), 73–84.

Malterud, K. (2001). Qualitative research: standards, challenges, and guidelines. *The Lancet*, *358*(9280), 483–488.

Marshall, M. N. (1996). Sampling for qualitative research. *Family Practice*, *13*(6), 522–526.

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

O'Reilly, M., & Parker, N. (2012). "Unsatisfactory Saturation": a critical exploration of the notion of saturated sample sizes in qualitative research. *Qualitative Research*, 1–8.

Patton, M. Q. (1990). *Qualitative evaluation and research methods .* SAGE Publications, inc.

Poole, M. S., Van de Ven, A. H., & Dooley, K. (2000). *Organizational Change and Innovation Processes*. Oxford: Oxford University Press .

Pratt, M. G. (2008). Fitting oval pegs into round holes tensions in evaluating and publishing qualitative research in top-tier North American journals. *Organizational Research Methods*, *11*(3), 481–509.

R Development Core Team. (2015). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Sandelowski, M. (1995). Sample size in qualitative research. *Research in Nursing & Health*, *18*(2), 179–183.

Shnayder, L., Van Rijnsoever, F. J., & Hekkert, M. P. (n.d.). Putting Your Money Where Your Mouth Is: Why sustainability reporting based on the triple bottom line can be misleading. *PLOS One*.

Suddaby, R. (2006). From the Editors: What Grounded Theory Is Not. *The Academy of Management Journal*, *49*(4), 633–642. doi:10.2307/20159789

Trotter II, R. T. (2012). Qualitative research sample design and sample size: Resolving and unresolved issues and inferential imperatives. *Preventive Medicine*, *55*(5), 398–400.

Tuckett, A. G. (2004). Qualitative research sampling: the very real complexities. *Nurse Researcher*, *12*(1), 47–61.

Walsh, I., Holton, J. A., Bailyn, L., Fernandez, W., Levina, N., & Glaser, B. (2015). What Grounded Theory Is … A Critically Reflective Conversation Among Scholars. *Organizational Research Methods* . doi:10.1177/1094428114565028

Yin, R. K. Case Study Research: Design and Methods (2009). Thousand Oaks, CA: Sage Publications Inc.

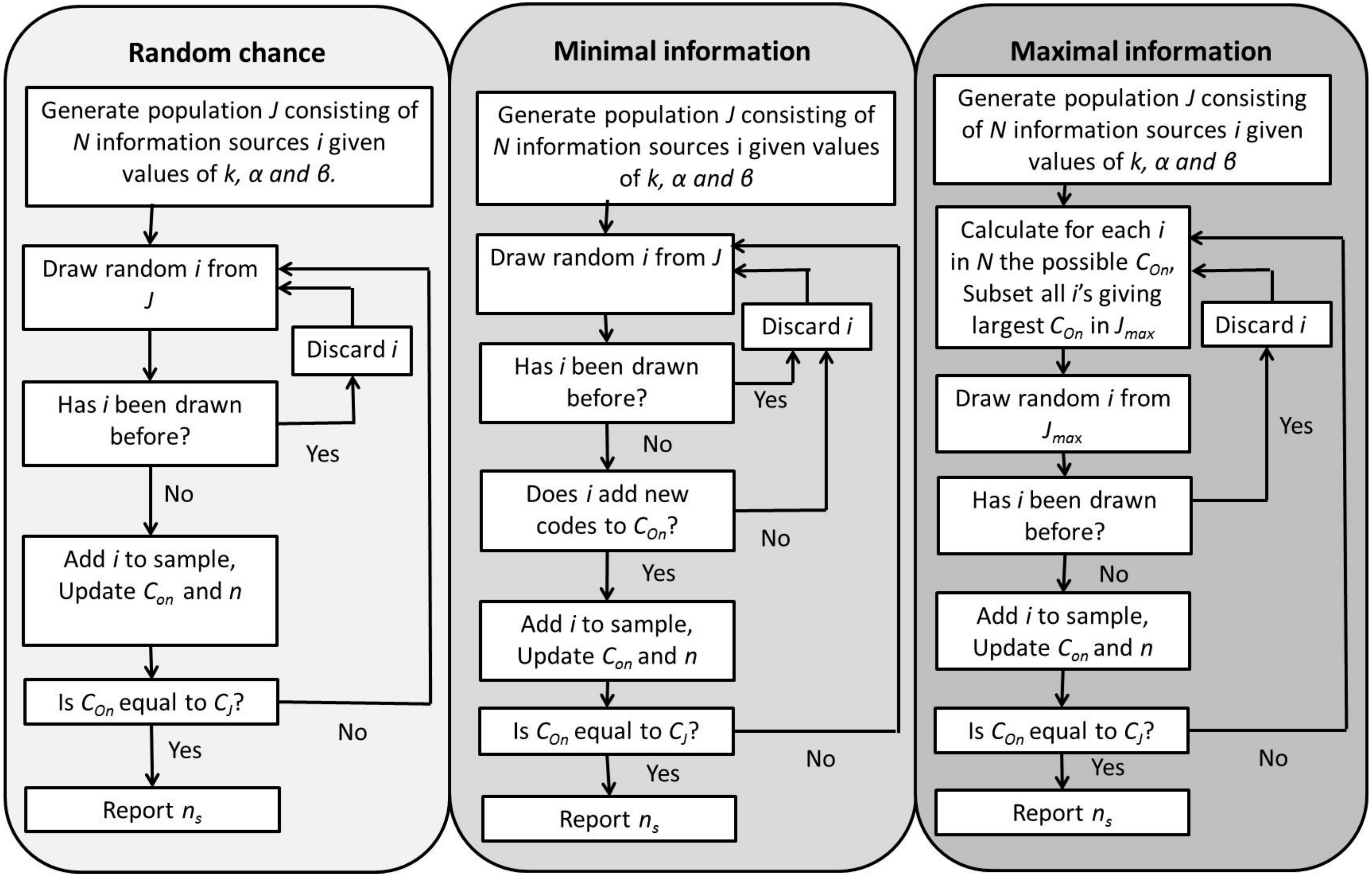
Random chance
Generate population J consisting of N information sources i given values of k, α and β.
Draw random i from J
Discard i
Has i been drawn before?
Yes
No
Add i to sample, Update C_{on} and n
Is C_{On} equal to C_J?
No
Yes
Report n_s
Minimal information
Generate population J consisting of N information sources i given values of k, α and β
Draw random i from J
Discard i
Has i been drawn before?
Yes
No
Does i add new codes to C_{On}?
No
Yes
Add i to sample, Update C_{on} and n
Is C_{On} equal to C_J?
No
Yes
Report n_s
Maximal information
Generate population J consisting of N information sources i given values of k, α and β
Calculate for each i in N the possible C_{On}, Subset all i's giving largest C_{On} in J_{max}
Discard i
Draw random i from J_{max}
Yes
Has i been drawn before?
No
Add i to sample, Update C_{on} and n
Is C_{On} equal to C_J?
Yes
No
Report n_s

Figure 1: a schematic overview of how the algorithms for each scenario operate.

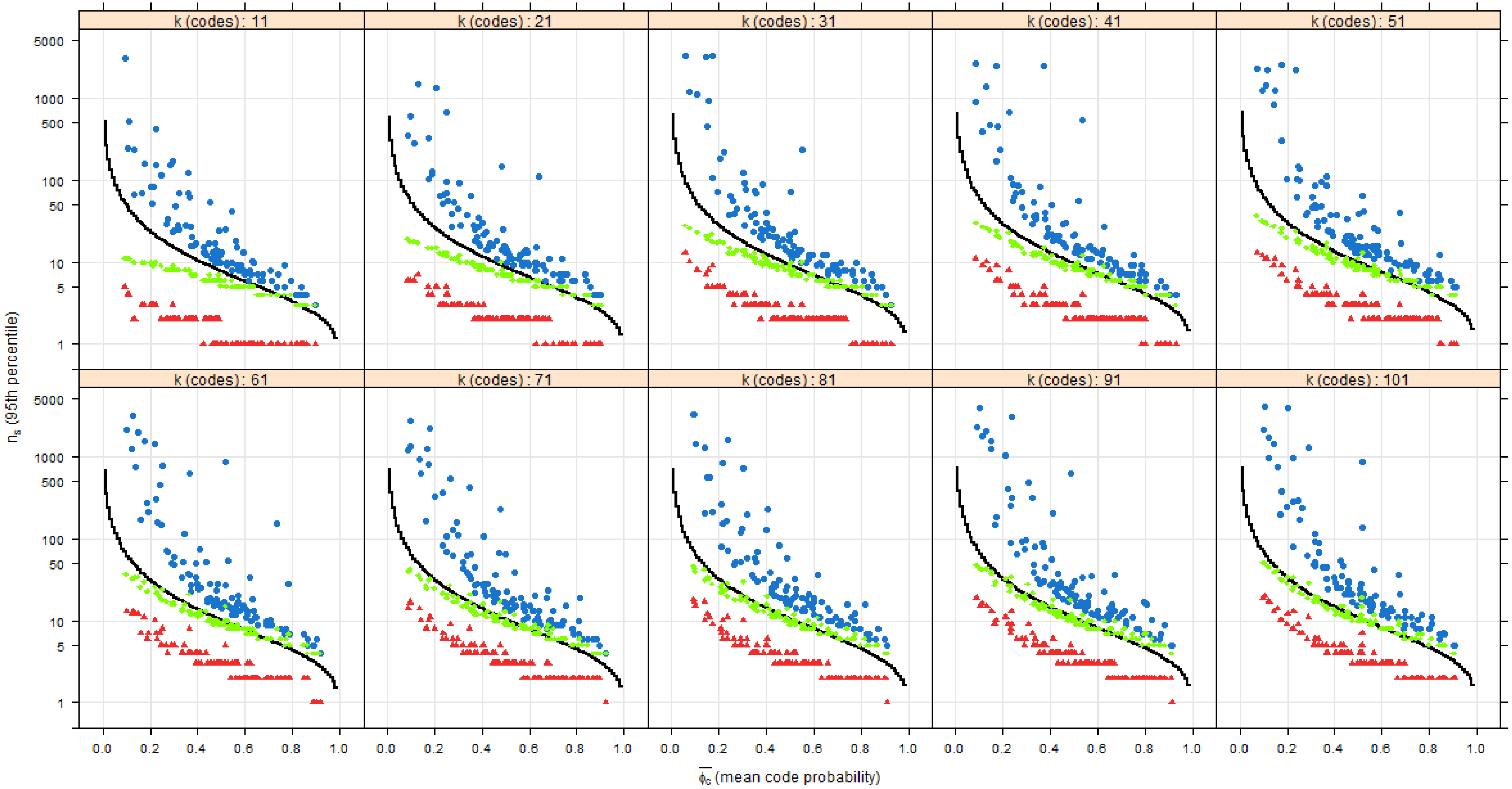

Figure 2: the 95th percentile of n_s (the number of sampling steps required to reach theoretical saturation) against $\overline{\Phi_c}$ (the mean probability of observing codes) for the values of k (the number of codes) between 11 and 101. Note that the y-axis is logarithmic. The solid black line indicates the calculated random chance's value of n based on *F11*. The blue dots represent random chance, the green diamonds represent minimal information, and the red triangles represent maximal information.

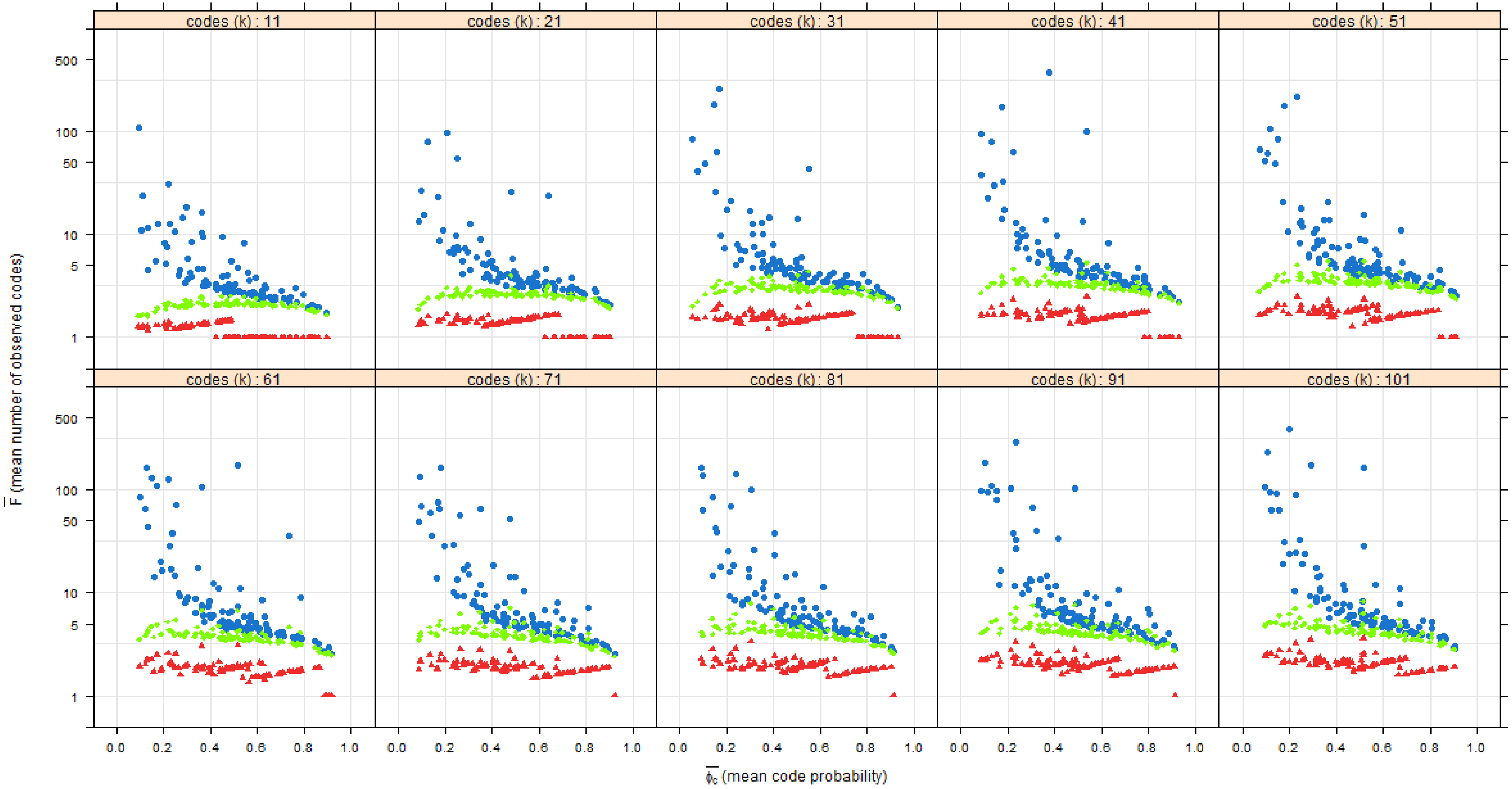

Figure 3: $\overline{F}$ (the mean number of observations per code) upon reaching theoretical saturation for each scenario against $\overline{\Phi_c}$ for the values of k between 11 and 101. The blue dots represent random chance, the green diamonds represent minimal information, and the red triangles represent maximal information.