



## ORIGINAL ARTICLE

# Predicting data saturation in qualitative surveys with mathematical models from ecological research

Viet-Thi Tran<sup>a,b,c,\*</sup>, Raphael Porcher<sup>b,c,d</sup>, Viet-Chi Tran<sup>e,f</sup>, Philippe Ravaud<sup>b,c,d,g</sup>

<sup>a</sup>Department of General Medicine, Paris Diderot University, 16 Rue Henri Huchard, 75018 Paris, France

<sup>b</sup>Centre de recherche en Épidémiologie et Statistiques (CRESS), INSERM U1153, Place du Parvis Notre Dame, 75004 Paris, France

<sup>c</sup>Centre d'Épidémiologie Clinique, Hôpital Hôtel-Dieu, Assistance Publique-Hôpitaux de Paris, 1 Place du Parvis Notre Dame, 75004 Paris, France

<sup>d</sup>Paris Descartes University, 12 Rue de l'École de Médecine, 75006 Paris, France

<sup>e</sup>Laboratoire Paul Painlevé—UMR CNRS 8524, Bâtiment M2, Cité Scientifique, 59655 Villeneuve-d'Ascq, France

<sup>f</sup>Université des Sciences et Technologies de Lille, Cité Scientifique, 59650 Villeneuve-d'Ascq, France

<sup>g</sup>Department of Epidemiology, Columbia University Mailman School of Public Health, 116th St & Broadway, New York, NY, USA

Accepted 3 October 2016; Published online xxxx

## Abstract

**Objective:** Sample size in surveys with open-ended questions relies on the principle of data saturation. Determining the point of data saturation is complex because researchers have information on only what they have found. The decision to stop data collection is solely dictated by the judgment and experience of researchers. In this article, we present how mathematical modeling may be used to describe and extrapolate the accumulation of themes during a study to help researchers determine the point of data saturation.

**Study Design and Setting:** The model considers a latent distribution of the probability of elicitation of all themes and infers the accumulation of themes as arising from a mixture of zero-truncated binomial distributions. We illustrate how the model could be used with data from a survey with open-ended questions on the burden of treatment involving 1,053 participants from 34 different countries and with various conditions. The performance of the model in predicting the number of themes to be found with the inclusion of new participants was investigated by Monte Carlo simulations. Then, we tested how the slope of the expected theme accumulation curve could be used as a stopping criterion for data collection in surveys with open-ended questions.

**Results:** By doubling the sample size after the inclusion of initial samples of 25 to 200 participants, the model reliably predicted the number of themes to be found. Mean estimation error ranged from 3% to 1% with simulated data and was <2% with data from the study of the burden of treatment. Sequentially calculating the slope of the expected theme accumulation curve for every five new participants included was a feasible approach to balance the benefits of including these new participants in the study. In our simulations, a stopping criterion based on a value of 0.05 for this slope allowed for identifying 97.5% of the themes while limiting the inclusion of participants eliciting nothing new in the study.

**Conclusion:** Mathematical models adapted from ecological research can accurately predict the point of data saturation in surveys with open-ended questions. © 2016 Elsevier Inc. All rights reserved.

**Keywords:** Sample size; Qualitative research; Data saturation; Open-ended questions; Surveys and questionnaires; Web-based questionnaires

## 1. Context

Surveys with open-ended questions are a simple design to explore the different aspects of a concept in a given

population [1]. This design is popular in many fields, including health research, social science, and marketing. For example, in health research, surveys may help identifying the topics that should be addressed in items of patient-reported outcomes [2]. The use of open-ended questions allows respondents to describe with nuance and detail how they perceive the concept under study. By reading and reflecting on participant answers, researchers can identify the meaningful variations and relationships of aspects of the concept, which allows for developing theories on how a particular phenomenon “works.”

Surveys with open-ended questions are related to qualitative research because they seek to describe the qualities of

Conflict of interest: None.

Funding: This study was funded by the French Health Ministry (PHRC AOM13127). Our team is supported by an academic grant from the program “Equipe espoir de la Recherche,” Fondation pour la Recherche Médicale, Paris, France (no. DEQ20101221475). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article.

\* Corresponding author. 1 place du Parvis Notre-Dame, 75181 Paris, France. Tel.: +33-1-42-34-89-87; fax: +33-1-42-34-87-90.

E-mail address: [thi.tran-viet@htd.aphp.fr](mailto:thi.tran-viet@htd.aphp.fr) (V.-T. Tran).

**What is new?****Key findings**

- Sample size in surveys with open-ended questions relies on the principle of data saturation. We used a mathematical model used in ecological research to describe and extrapolate the accumulation of themes during a survey using open-ended questions.
- Our model accurately predicted the point of data saturation in surveys with open-ended questions. We showed that the slope of the expected theme accumulation curve could be used as a stopping criterion in surveys using open-ended questions.

**What this adds to what was known?**

- Up to this date, sample size in surveys with open-ended questions was determined solely by the judgment and experience of researchers.
- Determining the number of themes or even describing exhaustively all themes is not an objective per se of qualitative research. However, our mathematical model can help researchers to estimate the number of participants to include and avoid small incomplete studies or needlessly large studies.

**What is the implication and what should change now?**

- We suggest the following analysis plan for determining sample size in surveys using open-ended questions:
  - Step 1: Invite a sample of 50 to 100 participants to respond and analyze their answers by using the researcher's preferred method.
  - Step 2: Organize findings as a matrix opposing observed themes and units of analysis.
  - Step 3: Use the model presented to compute the theme accumulation curve and the local slope of the curve at the point of data analysis.
  - Step 4: If the local slope of the theme accumulation curve is above the chosen stopping criterion, continue data collection and analysis and go back to step 2.

entities and phenomena that are not measured with numbers [3]. However, they also present differences from classic inquiry methods such as interviews or focus groups by the use of (1) structured questionnaires instead of free conversation, (2) a large sample of participants determined according to predefined criteria instead of purposeful

recruitment, and (3) linear data collection and analysis instead of circular and iterative processes.

Despite these differences, the purpose of surveys, similar to other qualitative inquiry methods, is the comprehensive and thorough description of the topic of interest. To that end, data collection and analysis continue to the point when additional input from new participants no longer changes the researchers' understanding of the concept. This is the point of data saturation [4,5]. Determining the number of participants to be included to obtain data saturation is one of the most frequent questions in qualitative research [6], but no transparent reproducible method has been developed to verify researchers' claims of having "seen nothing in newly sampled units or feeling comfortable that a theoretical category has been saturated" [7].

The problem for researchers in assessing the point of data saturation, that is, estimating the "true"—and unknown—number of themes about a given topic, is similar to that faced by ecological researchers when trying to estimate the number of species in an area [8]. In these studies, researchers cannot count or observe every possible animal and therefore use sampling methods to determine species richness. They use quadrats, lures, or traps in the study area; observe individuals captured; and enumerate the species found. From their empirical sample, researchers then use mathematical models to extrapolate the accumulation of species and determine the "true" species richness of an area. Such models have been used, for example, to determine the number of ant species in a tropical rain forest in Costa Rica [9].

The objective of this study was to present how mathematical modeling established in ecological research may be used to describe and extrapolate the accumulation of themes during a study using open-ended questions to help researchers to determine the point of data saturation. The specific aims are to (1) adapt the models established in ecological research to themes retrieved from health research to estimate the number of different themes that could be discovered in a survey with open-ended questions, (2) assess the reliability of this method by using both real data collected during a survey of the burden of treatment that used open-ended questions [10] and simulated data sets, (3) present a method to estimate the point in data collection when the inclusion of new participants is not likely to lead to the identification of new themes, and (4) propose an analysis plan for (health) research involving open-ended questions based on our results.

**2. Methods**

We used mathematical modeling to determine the point of data saturation in surveys using open-ended questions. It is important to note that the aim of our work was not to predict the themes, ideas, and meanings that patients may elicit on the topic of interest but rather to estimate how these new

ideas are discovered and accumulated across the whole sample of participants during a study.

### 2.1. Definitions

In the present study, we use the following definitions:

*Theme*: the atomic constituent of the phenomenon under study, in the context of the given study. To build theories, researchers analyze participants' inputs to identify and categorize themes into greater and more complex constructs by examining the relations between themes and/or participant characteristics.

*Unit of analysis*: the atomic unit on which the search for themes is conducted; can be a question, survey, participant interview, and so forth.

*Data collection*: all factors that may affect participant inputs, including the survey content, the formulation of questions, and contextual factors (e.g., environment or circumstances favoring participants' disclosure of private information).

*Elicitation of a theme*: the researcher identifying a given theme in a given unit of analysis. Thus, themes elicited (and their number) may vary depending on the researcher analyzing the data.

### 2.2. Presentation of the model

Inference on the process of theme accumulation using the themes already discovered is a complex task because researchers only have information on what they have found so far. In a given sample, the number of times a specific theme is elicited follows a binomial distribution with parameter the theme's prevalence, but this includes a number of naught for themes not yet found. Therefore, we cannot simply use a binomial model to predict the number of themes to be found, unless all themes are known. However, we can infer the total number of different themes, both known and unknown, to be found in a study under the following assumptions:

*Finite number of themes*: We assume that the number of themes in a study is finite and that this number is unknown to the researchers.

*Constant probability of theme detection*: We assume that methods for data collection, analysis, and factors influencing them are fixed. The probability of detecting a theme does not vary during the study and is identical for each unit of analysis.

*Independence of themes*: We assume that themes elicited in a given unit of analysis are independent of each other.

In a method developed for ecological research, Mao et al. [11] considered a latent distribution for the probability of elicitation of all themes (elicited or not) and used it to model the number of themes elicited in a study

as arising from a mixture of zero-truncated binomial distributions. Estimation of this latent distribution is possible with nonparametric maximum likelihood methods and expectation maximization (EM) algorithms. By using the estimated latent distribution, one can extrapolate the theme accumulation curve beyond the number of units of analysis already sampled and thus estimate the number of themes that could be found when adding more units of analysis in the sample. Confidence intervals are obtained by bootstrapping. Mathematical elements are presented in Appendix 1 at [www.jclinepi.com](http://www.jclinepi.com). A full demonstration is presented in the original article by Mao et al. [11].

This model uses as data a table crossing participants and themes elicited. Such tables can be easily obtained by using common qualitative analysis software such as QSR International's NVivo Software (<http://www.qsrinternational.com/support/faqs/how-do-i-cite-nvivo-10-nvivo-9-or-nvivo-8-in-my-wo>) with the framework matrix tool.

### 2.3. Performance of the model

We evaluated the performance of the model to predict the number of themes that could be obtained by adding more units of analyses. For that, we calculated the predicted number of themes that could be obtained by doubling the sample size after the initial inclusion of 25, 50, 100, or 200 units of analysis and compared it with the number of themes actually elicited after the addition of these new units.

### 2.4. Use of a stopping criterion to determine the point of data saturation

An exhaustive description of a phenomenon with the identification of all possible themes might require very large samples. Thus, the question may not be "How many units should I collect to obtain data saturation?" but rather "At which point in data collection does the cost of including a new unit of analysis exceed the expected gain in information?" A simple method to determine the expected gain in information with the addition of new data is the estimation of the local slope of the expected theme accumulation curve. The slope represents the expected number of themes to be found with the inclusion of "n" new units of analysis: for example, a slope of 0.05 would indicate that one new theme is expected to be found after the inclusion of 20 new units of analysis.

We propose that researchers could use the slope of the theme accumulation curve as a stopping criterion for recruitment. To choose a slope value that strikes a balance between the probability of missing some themes and sample size, we sequentially calculated the value after every 5 units included in the study. Then, we assessed the number of units recruited and the number of themes discovered if

we had used slope values of 0.20, 0.10, 0.05, or 0.01 as stopping criteria.

## 2.5. Material

To illustrate how our model could be used and its performance, we used (1) empirical data gathered in a study on the burden of treatment [10] and (2) by Monte Carlo simulations with data sets mimicking 1,000 surveys with open-ended questions.

### 2.5.1. Study of the burden of treatment

In the study of the burden of treatment, 1,053 English-, Spanish-, and French-speaking patients with at least one chronic condition answered 15 open-ended questions in an Internet survey about the difficulties they had with their health care. The mean participant age was 46 (standard deviation = 14) years, and they resided mainly in France (64%), the United States (13%), Canada (6.3%), the United Kingdom (5.3%), Spain (3.2%), and Australia (2.8%). Self-reported main chronic conditions ranged from rheumatologic diseases (33%) to cancers (8%).

Participants' answers were the unit of analysis and were examined by content analysis. In a first step, for the first 200 responses in French and English, two investigators independently identified "in vivo codes": literal terms used by participants to describe their burden of treatment. During meetings, the investigators reached consensus on the initial codes and grouped them into an initial set of themes. For example, "I will have to take medication for the rest of my life, there aren't holidays for treatment" was coded as a theme entitled "Treatment is for whole life." In a second step, this initial set of themes was used for analysis of the remaining responses: each participant's response was read by two investigators (at least one researcher native in the given language), who independently assigned data segments to each theme. During meetings, the investigators compared their analyses and reached consensus on coding. Whenever a new idea emerged, researchers discussed the idea, thereby refining and enriching the list of themes. As a result, we could describe the different themes each participant mentioned and thus create a table opposing participants and themes elicited.

Answers to open-ended questions formed an overall corpus of 408,625 words. The median (Q1–Q3) length of patient answers was 298 (129–526) words globally (maximum 2,699). Content analysis of participant answers led to the identification of 123 different themes. Themes related to (1) tasks imposed on patients by their diseases and by their health care system (e.g., medication management, lifestyle changes, follow-up); (2) structural factors (e.g., access to health care resources) and personal, situational, and financial factors that aggravated the burden of treatment; and (3) the consequences of the burden of disease (e.g., poor adherence to treatments, financial burden, impact on professional, family, and social life).

All themes had been mentioned at least once after the 681st patient had been included. Themes in our burden of treatment study had low intercorrelation within a patient; the mean correlation between themes was  $r = 0.03$  (range  $-0.003$  to  $0.07$ ). Results of the qualitative study on the burden of treatment are described in another study [10].

### 2.5.2. Simulated data sets

To further document the properties of the model, we used 2 groups of simulated data sets. Each simulated data set mimicked a survey with open-ended questions, with 1,000 units of analysis, eliciting a total of 120 themes.

The first group of data sets tested the reliability of the model to accurately predict the number of themes to be found and the efficacy of stopping criteria based on the slope of the theme accumulation curve. It consisted of 1,000 data sets in which the probability of eliciting the themes and correlations were similar to those found in our burden of treatment study.

The second group of data sets tested the robustness of the model in cases of higher correlation between themes. It consisted of 1,000 data sets in which the themes elicited showed intercorrelation. The mean correlation between themes was  $r = 0.29$  (range  $0.17$ – $0.34$ ), approximately 10 times higher than in our burden of treatment study.

## 3. Results

### 3.1. Performance of the model

In both our study of the burden of treatment and the simulated data sets, the model reliably predicted the number of themes to be found by doubling the sample size of a study. In our study of the burden of treatment, the prediction errors were  $<2\%$  (difference between expected and observed number of themes were at most 2 of 123 themes) with initial samples of 25, 50, 100, and 200 participants (Table 1 and Fig. 1).

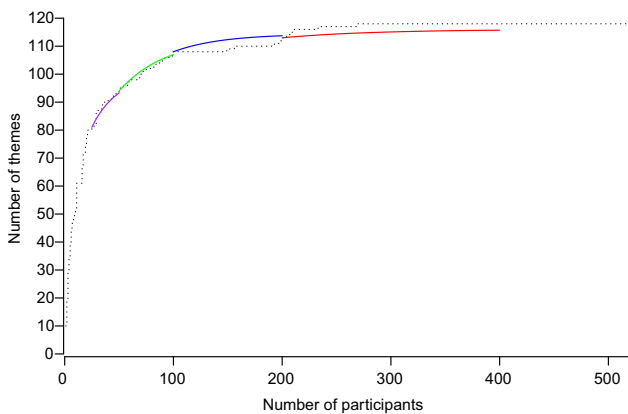
The excellent predictive capability of the model was confirmed with the first group of simulated data sets. The median differences between expected and observed number of themes after doubling the sample size ranged from 3.9 (interquartile range [IQR] =  $-2.9$  to  $16.3$ ) for an initial

**Table 1.** Expected number of themes that could be discovered by doubling the sample size after inclusion of 25, 50, 100, and 200 participants in the burden of treatment study

Initial sample size	Expected no. of themes by doubling the sample size	Actual no of themes by doubling the sample size
25	93 (81–116)	94
50	107 (92–130)	108
100	114 (99–139)	113
200	116 (102–142)	118

Data are no. (95% CI). Confidence intervals obtained with a bootstrap method.





**Fig. 1.** Number of themes that could be elicited by doubling the sample size with initial samples of 25, 50, 100, and 200 participants in the burden of treatment study. Dotted line represents the observed theme accumulation curve during the study. Colored curves represent expected number of themes calculated by using the mathematical model. Purple, blue, red, and green lines represent extrapolation of data obtained with 25, 50, 100, and 200 participants, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

sample of 25 units to  $-1.5$  ( $-2.8$  to  $-0.24$ ) for an initial sample of 200 units of analysis (Fig. 2).

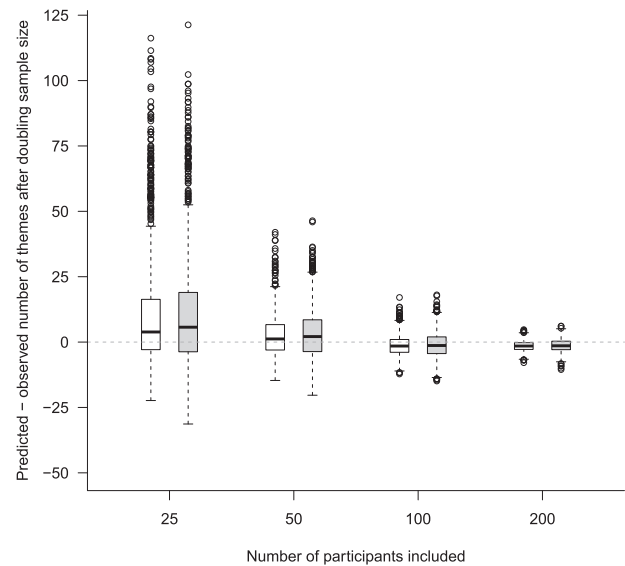
The model was also robust. With the second group of simulated data sets, with higher correlation between themes, the median difference between expected and observed number of themes after doubling the sample size ranged from 5.7 (IQR =  $-3.7$  to 19.0) for an initial sample of 25 units to  $-1.4$  ( $-2.9$  to 0.34) for an initial sample of 200 units of analysis (Fig. 2).

### 3.2. Use of a stopping criterion to determine point of data saturation

Use of the slope of the expected theme accumulation curve was a feasible approach to determine the cost–benefit ratio of adding new units of analysis in a survey with open-ended questions and thus could be used as stopping criterion for recruitment.

In our study of the burden of treatment, the slope of the theme accumulation curve after the inclusion of 100 participants was 0.17, so approximately eight new participants would be required to discover a new theme. After the inclusion of 200 participants, the slope was 0.03 (i.e., approximately 29 new participants would be required to discover a new theme; Table 2; Fig. 3).

With the simulated data sets, we compared the number of themes identified and sample sizes required when using the slope of the theme accumulation curve as a stopping criterion for recruitment. A slope of 0.20 allowed for identifying a mean of 113/120 themes (with 0.5% simulations achieving “true” data saturation; i.e., 100% of themes identified) with a mean of 149 units of analysis (Fig. 4). Use of a slope of 0.05 as a stopping criterion allowed for



**Fig. 2.** Difference between expected and observed number of themes elicited after doubling the sample size using the simulated data. White box plots represent simulations for performance (probability of theme elicitation is similar to that in the burden of treatment study). Gray box plots represent simulations for robustness with high theme correlations. Horizontal lines are median; outer edges are inter-quartile range; whiskers are minimum, maximum values.

identifying a mean of 117/120 themes (with 6% of simulations achieving “true” data saturation) with a mean of 230 units of analysis. Finally, use of a slope of 0.01 allowed for identifying a mean of 118/120 themes (with 30% of simulations achieving “true” data saturation) with a mean of 348 units of analysis. Obviously, using smaller values of the slope as a stopping criterion led to the identification of more themes but at the cost of more units of analysis.

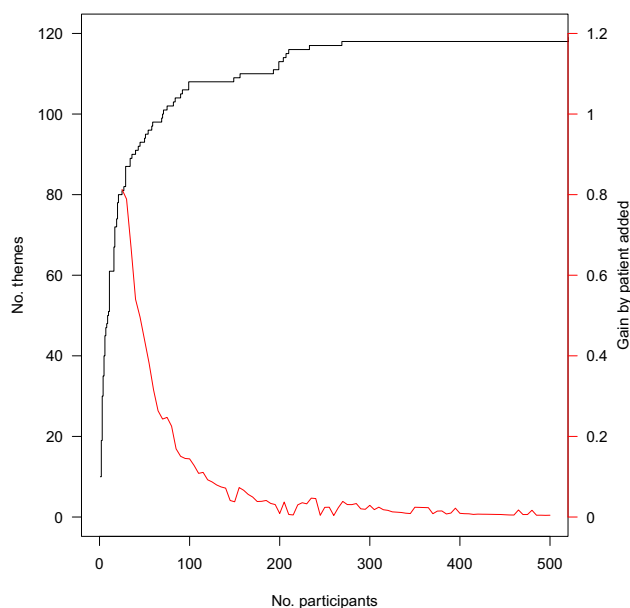
The presence of a high correlation between themes led to the recruitment of more units, but results were approximately comparable to the previous simulations. For instance, using a slope of 0.05 as a stopping criterion allowed for identifying a mean of 116/120 themes (with 7% simulations achieving “true” data saturation) with a mean of 229 units (Appendix 2 at [www.jclinepi.com](http://www.jclinepi.com)).

## 4. Discussion

In this study, we showed that models used in ecology to determine species richness could help with qualitative

**Table 2.** Estimated number of new units of analysis required to discover an additional theme after the inclusion of 25, 50, 100, 200, and 400 participants in the burden of treatment study

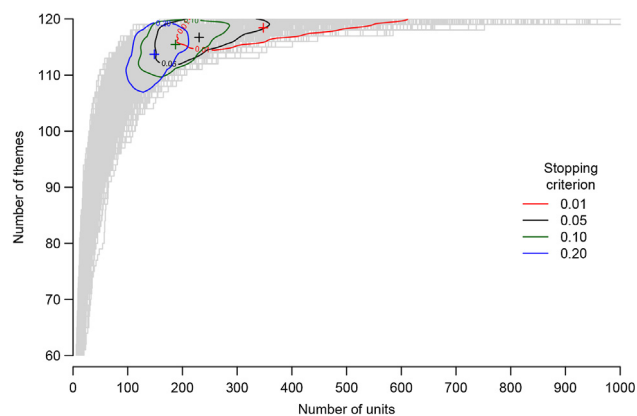
No. of participants already recruited	Estimated no. of patients required to discover an additional theme
25	1
50	2
100	8
200	28
400	108



**Fig. 3.** Theme accumulation curve (black curve) and expected gain in number of themes elicited for inclusion of a new participant (red curve) in the burden of treatment study. The expected gain corresponds to the local slope of the expected theme accumulation curve, calculated sequentially after the inclusion of each participant. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

research involving surveys with open-ended questions to predict what themes will be discovered with the inclusion of more units of analysis. Determining when to stop data collection is a thorny question asked by both novice and experienced researchers in qualitative research [6]. However, there is a surprising paucity of explicit discussion of this basic issue in textbooks and articles [5,6]. Authors usually explain that the point of data saturation appears evident to researchers during the iterative cycles of data collection and analysis when they know their data, understand it intimately, and “get inside it” [12]. However, such process is opaque and a growing concern is researchers stopping data collection not because the collection of new data does not shed any further light on the issue under investigation but rather for logistic or financial reasons [13,14].

In this study, we do not pretend to have solved the problem, but we offer a reliable and reproducible solution for research involving surveys with open-ended questions to estimate the point of data saturation. Because the model involves data collected from an empirical sample, it takes into account both the reliability of the survey instrument (i.e., the ability for the open-ended questions to provoke respondents’ answers) and the way data are analyzed (i.e., the granularity and level of detail desired by researchers). We showed that our model accurately predicted the number of themes that could be discovered by doubling the sample size. Precision of the estimation grew with the size of the data set used for calibration of the model. Although prediction was possible with sample sizes as small as 25



**Fig. 4.** Number of themes discovered and number of patients included when using a local slope of the theme accumulation curve as a stopping criterion with the simulated data used for performance. Each gray line represents a single simulated data set of 1,000 units of analysis eliciting a total of 120 different themes. Blue, green, black, and red contours show where 90% of simulations are stopped when using a stopping criterion of 0.20, 0.10, 0.05, and 0.01, respectively, based on the slope of the theme accumulation curve. Colored points show the mean number of themes discovered and mean number of patients recruited for each stopping criterion tested. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

participants; in some extreme cases (<5% cases), it was not accurate. In our simulated data sets, an initial sample size of 50 units of analysis ensured accurate prediction of the number of themes to be elicited with twice as many units of analysis for all possible simulations.

To help qualitative researchers decide when to stop data collection, we tested the performance of different values of a stopping criterion based on the slope of the theme accumulation curve. In our opinion, a cutoff of 0.05 (i.e., 20 more participants are required to identify a new theme) exhibited a good balance between number of themes discovered and number of units included.

Theme accumulation curves are not a new idea in qualitative inquiry methods. Guest et al. [5] drew charts of the identification of new themes over the course of their studies. However, their results were limited to a single case and could not be generalized to other settings, populations, or study aims. Francis et al. also drew a theme accumulation curve and suggested defining a stopping criterion of three interviews leading to no identification of new data [15]. However, in a previous work, we showed that this stopping rule was not reliable and did not approach the point of data saturation [16]. Our approach using mathematical models is original in the field and offers both transparency and reproducibility.

From this study and our experience, we suggest the following analysis plan for research involving surveys with open-ended questions (Box 1):

*Step 1:* Instead of sending a large number of questionnaires to the population of interest, invite a sample of

### Box 1 Suggested analysis plan for research involving surveys with open-ended questions

Step 1: Invite a sample of 50–100 participants to respond and analyze their answers by using the researcher's preferred method.

Step 2: Organize findings as a matrix opposing observed themes and units of analysis.

Step 3: Use the model presented to compute the theme accumulation curve and the local slope of the curve at the point of data analysis.

Step 4: If the local slope of the theme accumulation curve is above the chosen stopping criterion, continue data collection and analysis and go back to step 2.

50 to 100 participants to respond and analyze their answers by using the researcher's preferred method.

Step 2: Organize findings as a matrix opposing observed themes and units of analysis.

Step 3: Use the model presented to compute the theme accumulation curve and the local slope of the curve at the point of data analysis.

Step 4: If the local slope of the theme accumulation curve is greater than the chosen stopping criterion, continue data collection and analysis and go back to step 2.

This study has some limitations. First, the model supposes that data collection and analysis methods remain fixed during the study. This situation is often not true because one of the central features of qualitative analysis (whatever the paradigm and method) is the researcher "getting inside the data" and using insights from what he read before to help better understand new inputs. However, despite this assumption, our model showed excellent prediction capabilities for our study of the burden of treatment and in the simulated data sets. Second, one assumption of the model is the independence of themes elicited by a given unit of analysis. Although this assumption was verified in our study of the burden of treatment, other studies, in different contexts and with different topics, could present greater interthemes correlation, which may bias predictions from the model. Finally, other methods from ecological research could be used to model data collected and analysis from surveys with open-ended questions. Although results with our model were excellent, different models [17] may have better (or different) properties.

Finally, we acknowledge that the number of themes, or even the exhaustive description of all themes, does not mean that researchers will be able to understand the relationships between them and is not an objective per se of qualitative research. We do not intend to fan the

qualitative/quantitative debate: researchers should always try to consider the best methods for conducting their research. In this particular case, we believe that the use of mathematics could help researchers estimate the number of participants to include for eliciting themes and thus avoid small incomplete studies or needlessly large studies leading to waste of time and effort for researchers and participants.

## 5. Conclusions

In surveys with open-ended questions, the point of data saturation and number of participants to include can be estimated with mathematical models from ecological research.

## Acknowledgments

The authors thank Laura Smales (BioMedEditing) for editing.

Authors' contributions: V.-T.T., R.P., V.-C.T., and P.R. conceived and designed the experiments. V.-T.T. and R.P. analyzed data. V.-T.T. wrote the first draft of the article. V.-T.T., R.P., V.-C.T., and P.R. contributed to the writing of the article. V.-T.T., R.P., V.-C.T., and P.R. met ICMJE criteria for authorship. V.-T.T., R.P., V.-C.T., and P.R. agreed with article results and conclusions. P.R. is the guarantor, had full access to the data in the study, and takes responsibility for the integrity of the data and the accuracy of the data analysis.

## References

- [1] Jansen H. The logic of qualitative survey research and its position in the field of social research methods. *Forum Qual Social Res* 2010; 11(2). Art. 11. Available at <http://www.qualitative-research.net/index.php/fqs/article/view/1450/2946>. Accessed November 09, 2016.
- [2] Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;60:34–42.
- [3] Denzin N, Lincoln Y. The discipline and practice of qualitative research. In: Denzin N, Lincoln Y, editors. *The SAGE Handbook of Qualitative Research*. Thousand Oaks, CA: SAGE; 2011.
- [4] Glaser B, Strauss A. *The discovery of grounded theory: strategies for qualitative research*. Chicago, IL: Aldine Publishing Company; 1967.
- [5] Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. *Field Methods* 2006; 18:59–82.
- [6] Baker S, Edwards R. How many qualitative interviews are enough? Expert voices and early career reflections on sampling and cases in qualitative research. NCRM National Centre for Research methods; Southampton; 2012.
- [7] Sandelowski M. Sample size in qualitative research. *Res Nurs Health* 1995;18:179–83.
- [8] Ugland K, Gray J, Ellingsen K. The species accumulation curve and estimation of species richness. *J Anim Ecol* 2003;72: 888–97.
- [9] Longino J, Coddington J, Colwell RK. The ant fauna of a tropical rain forest: estimating species richness three different ways. *Ecology* 2002;83(3):689–702.

- [10] Tran VT, Barnes C, Montori VM, Falissard B, Ravaud P. Taxonomy of the burden of treatment: a multicountry Web-based qualitative study of patients with chronic conditions. *BMC Med* 2015;13:115.
- [11] Mao CX, Colwell RK, Chang J. Estimating the species accumulation curve using mixtures. *Biometrics* 2005;61:433–41.
- [12] Morse JM. “Data were saturated...”. *Qual Health Res* 2015;25:587–8.
- [13] Mason M. Sample size and saturation in PhD studies using qualitative interviews. *Forum Qual Social Res* 2010;11(3). Art. 8. Available at <http://www.qualitative-research.net/index.php/fqs/article/view/Article/1428/3027>. Accessed November 09, 2016.
- [14] Charmaz K. *Constructing grounded theory: a practical guide through qualitative analysis*. Thousand Oaks, CA: SAGE; 2006.
- [15] Francis JJ, Johnston M, Robertson C, Glidewell L, Entwistle V, Eccles MP, et al. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychol Health* 2010;25(10):1229–45.
- [16] Tran VT, Porcher R, Falissard B, Ravaud P. Point of data saturation was assessed using resampling methods in a survey with open-ended questions. *J Clin Epidemiol* 2016. <http://dx.doi.org/10.1016/j.jclinepi.2016.07.014>. Available at <https://www.ncbi.nlm.nih.gov/pubmed/27492788>.
- [17] Chao A, Colwell RK, Lin CW, Gotelli NJ. Sufficient sampling for asymptotic minimum species richness estimators. *Ecology* 2009; 90(4):1125–33.



**Appendix 1.**

**1. Presentation of the model**

We present here the model to extrapolate species accumulation curves developed by Mao et al. [1]. We used it to estimate the “true” number of themes that could be discovered in the target population.

We note  $S$  the “true” number of relevant themes that could be discovered in the target population.

We note  $h$  the number of units of analysis in the empirical sample already recruited by researchers.

We note  $G$  the number of groups of themes with the same probability to be elicited.

For the  $k$ th group, we define  $\psi_k$  the common probability of elicitation and  $\pi_k$  the relative group size (i.e., number of themes present in that group divided by the total number of  $S$ ).

$$\tau(h) = S \sum_{k=1}^G \pi_k [1 - (1 - \psi_k)^h] \tag{1}$$

Extrapolation of  $\tau(h)$  with  $h > H$  where  $H$  is the number of units of analysis in the empirical sample is not simple. Mao et al. describe a method to extrapolate  $\tau(h)$  for  $h > H$  using a likelihood-based method [1] under the assumption that  $G < H/2$ . This method relies on the development of a function  $\theta(h)$ :

$$\tau(h) = \tau(H)\theta(h) \tag{2}$$

As the expected number of themes  $\tau(h)$  results from a mixture of binomial distributions of  $G$  groups, we can define the mixing weight  $\omega_k$  for the  $k$ th group using the relative group size  $\pi_k$  and the common probability of elicitation of the themes  $\psi_k$  in that group:

$$\omega_k = \frac{\pi_k [1 - (1 - \psi_k)^H]}{\sum_{m=1}^G \pi_m [1 - (1 - \psi_m)^H]} \tag{3}$$

And:

$$\theta(h) = 1 + \sum_{k=1}^G \omega_k \frac{(1 - \psi_k)^H - (1 - \psi_k)^h}{1 - (1 - \psi_k)^H} \tag{4}$$

Estimation of parameters parameters  $\psi_k$  and  $\omega_k$  can be done by maximizing the log conditional likelihood

$$L = l(\{\omega_k, \psi_k\}_{k=1}^G) \tag{5}$$

of the empirical counts  $S_1, S_2, \dots, S_H$  given the observed number of themes  $S_{obs}$ .

For a given number of incidence groups  $G$ , an EM algorithm can be used to maximize the log likelihood  $L$ , yielding a set of estimators  $\psi_k$  and  $\omega_k$  that are fitted for  $G$  groups. To determine an optimal number of groups  $G$  that maximize the log likelihood without adding too much complexity, Mao et al. suggest the use of a number of groups  $G$  that minimize the Akaike information criterion (AIC):  $AIC = 2(2G - 1) - 2l(\{\omega_k, \psi_k\}_{k=1}^G)$ .

Using estimates of  $\psi_k$  and  $\omega_k$  obtained using the EM algorithm, it is then possible to determine  $\tau(h)$  for  $h > H$ . Confidence intervals were obtained computing 1,000 bootstraps from the likelihood counts  $S_1, S_2, \dots, S_H$  using a random  $\check{S}_{obs}$  from a Poisson distribution  $Poi(S_{obs})$ .

**2. Use of a stopping criterion based on the slope of the theme accumulation curve**

To determine the expected gain in information for addition of a new data, we estimated the slope of the expected theme accumulation curve  $V_i(h)$  between the  $h$ th and the  $h + i$ th unit, after the inclusion of  $h$  units:

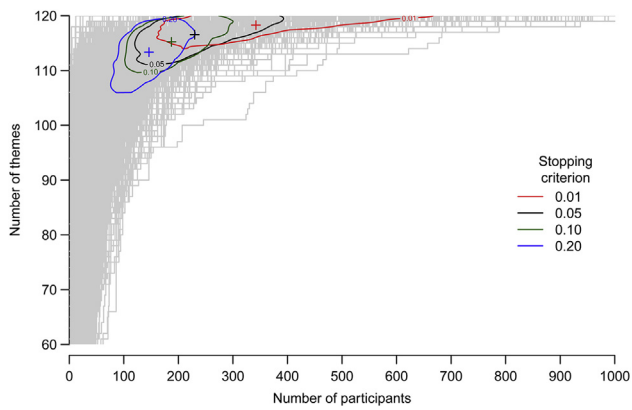
$$V_i(h) = \frac{\tau(h+i) - \tau(h)}{i} \tag{6}$$

**Reference**

[1] Mao CX, Colwell RK, Chang J. Estimating the species accumulation curve using mixtures. *Biometrics* 2005;61:433–41.

8.e2

V.-T. Tran et al. / Journal of Clinical Epidemiology xx (2016) 1–8



**Appendix 2.** Number of themes discovered and number of patients included when using a local slope of the theme accumulation curve as a stopping criterion with the simulated data for robustness. Each gray line represents a single simulated data set of 1,000 units of analysis eliciting a total of 120 different themes. Blue, green, black, and red contours show where 90% of simulations are stopped when using a stopping criterion of 0.20, 0.10, 0.05, and 0.01, respectively, based on the slope of the theme accumulation curve. Colored points show the mean number of themes discovered and mean number of patients recruited for each stopping criterion tested.