The Role of Data Simulation in Quantitative Ethnography

Zachari Swiecki[1(✉)] and Brendan Eagan[2]

[1] Monash University, Clayton, VIC 3800, Australia
zach.swiecki@monash.edu
[2] University of Wisconsin-Madison, Madison, WI 53711, USA

Abstract. Data simulations are powerful analytic tools that give researchers a great degree of control over data collection and experimental design. Despite these advantages, data simulations have not yet received the same amount of use as other techniques within the context of quantitative ethnography. In this paper, we explore the reasons for this and use examples of recent work to argue that data simulations can—and already do—play an important role in quantitative ethnography.

Keywords: data simulation · quantitative ethnographic methods · epistemic network analysis

1 Introduction

Data simulations are powerful analytical tools. Like statistical models, they can quantitatively represent phenomena that we observe in the world. Unlike statistical models, they are used to generate hypothetical data rather than predictions or inferences from real data. In turn, they afford researchers a high degree of control over parts of a study that are typically arduous, complex, and time consuming—things like data collection and experimental design.

Despite these advantages, data simulation has not been widely adopted as a quantitative ethnographic technique. Understanding why is not particularly difficult. Setting aside the training and experience required to develop data simulations, quantitative ethnography (QE) depends on the alignment between observed phenomena, qualitative claims, and quantitative warrants. More specifically, it depends on the alignment between *real* data about *real* events and qualitative and quantitative interpretations. Because data simulations by definition do not produce real data, they seem to have no place in QE.

In this paper, we argue that even though data simulations generate and operate on hypothetical data, they can—and already do—play a useful role in quantitative ethnographic analyses and the development of QE tools and methods.

2 Background

Before we describe the role of data simulation in QE, it will be useful to review some of the finer points of the QE process [14, 16] (Fig. 1). While these points are crucial to

C. Damşa and A. Barany (Eds.): ICQE 2022, CCIS 1785, pp. 87–100, 2023.
https://doi.org/10.1007/978-3-031-31726-2_7

QE, they may nonetheless be obfuscated in a typical analysis given the complexity of many QE techniques.

QE is fundamentally a process for providing quantitative warrants for qualitative claims. These claims are made in terms of the [D]iscourse of some culture—that is, the ways in which members of that particular culture act, talk, think, believe, value, solve problems, and so on. To make these claims, researchers observe the actual things members of the culture say and do and record them in some way—they observe the [d]iscourse of the culture and record their observations as some form of data (field notes, audio/video recordings, interaction logs, and so on).

The translation of [d]iscourse to data is the first of many simplifications of the [D]iscourse of a culture that are necessary to conduct any QE analysis. Of all the possible things researchers could observe members of a culture doing, they observe and record some subset of those things as data. And as happens in any human endeavour, they may make errors.

Prior to analyzing their data, researchers often make another simplification—they translate their data to some other—usually machine-readable—format. For example, field notes may be typed up or audio may be transcribed. Here again, errors may occur; notes may be mistyped, audio mistranscribed.

Using their recorded data, researchers attempt to understand the relationships between particular themes, ideas, or actions that members of a particular culture use to understand and operate on the world. In other words, they look for evidence of [C]odes in their data. Their evidence comes in the form of [c]odes, identifiable pieces of data that indicate the presence of [C]odes.

The act of coding is an act of pointing; it is a way of saying that some identifiable piece of data is representative of a higher-level concept [16]. To warrant these acts of pointing, QE researchers marshal a collection of qualitative and quantitative evidence. After qualitatively examining the data, they develop a *codebook* that describes examples of the links between the higher level concepts they are investigating ([C]odes) and how those concepts are instanced in the actual data they have ([c]odes).

Using a codebook, two or more raters apply it to the data—otherwise known as *coding*—annotating segments of the data for the presence or absence of the [C]ode. They then compare their ratings using inter-rater reliability (IRR) metrics such as Cohen's kappa and Shaffer's rho to demonstrate that these decisions can be reliably applied to the data. Of course, researchers may also develop automated classifiers to identify [C]odes using techniques such as regular expressions and evaluate them in a similar way.

The process of coding data is, of course, another simplification with the potential to introduce error. Quantitative metrics like kappa and rho are a way of controlling for these kinds of errors, a way of measuring the error and setting thresholds for how much of it they are willing to tolerate in the analysis.

Once codes have been identified and their relationship to [C]odes warranted, the next step is typically to identify relationships among the [c]odes that are salient to the purpose of the analysis. To warrant that these relationships, or connections, constitute systematic patterns in the data and not simply one-off or random occurrences—that is, to warrant theoretical saturation—researchers represent the relationships among [c]odes using statistical models such as epistemic network analysis (ENA) [15]. They then test

whether a value derived from the sample of data is representative of what that value would be if calculated from the larger population of data that they might have collected about the same participants under similar conditions [17]. Here we denote the value from the sample as a [p]arameter and the value from the population as a [P]arameter.

Having found this quantitative warrant, researchers now have evidence that the relationships among [c]odes that they observed in their data are representative of the relationships among the corresponding [C]odes that shape the [D]iscourse of the culture they are studying. In other words, their qualitative claims are a systematic property of that culture's [D]iscourse. Crucially, however, the QE researcher's task is not complete until they re-examine these claims in terms of the actual data they have collected. That is, after the sometimes long, complicated, and reductive task of operationalizing qualitative claims in quantitative terms, researchers should check that their quantitative representations are aligned with—or not contradicted by—the actual observations they have made. In other words, they need to *close the interpretive loop.*

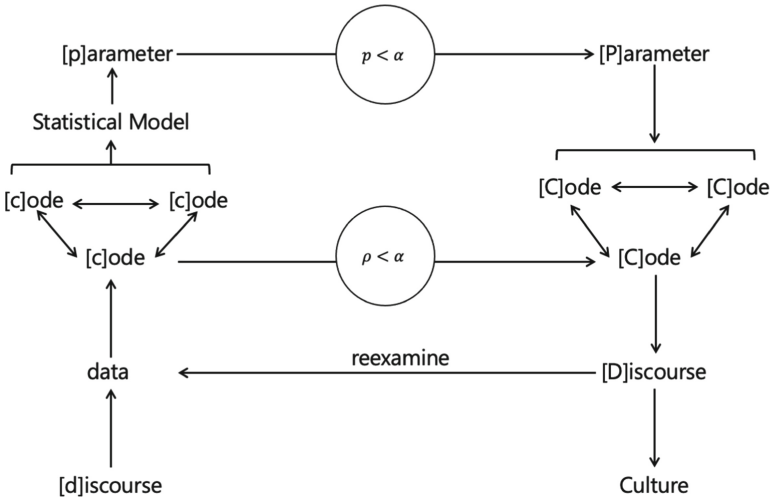

Fig. 1. The QE process. Adapted from [14]

The description above highlights two important features of the QE process. First—like any form of analysis—QE requires simplifications of the phenomena researchers wish to investigate. The things we observe members of a particular culture say and do we record as data; we categorise the kinds of things that members of that culture do by pointing to specific pieces of data; we look for connections among these categories by identifying relationships among these pieces of data. There is always the danger of oversimplification and error. Observations can be misrecorded; parties may not agree on whether some pieces of data actually correspond to categorizations of cultural activities; and identified connections may be meaningless. Second, the QE process hinges on the alignments between qualitative claims, quantitative representations, and real—that is, actually observed—data.

3 Data Simulation

As the term "data simulation" suggests, this technique does not traffic in real data. To highlight this feature, Gilbert and Troitzsch [9] argue that data simulation differs importantly from traditional statistical modeling (Fig. 2). In the latter, researchers have some real-world target that they want to understand. Their aim is to create a model of the target that is easier to study than the target itself. To do so, they collect data and develop a model (e.g., a set of regression equations) that abstracts salient features of the target. This model includes some parameters (e.g., beta coefficients) whose magnitudes are determined by fitting the model to the data on hand. Finally, they test whether the model generates predictions that are sufficiently similar to the collected data (e.g., using a coefficient of determination) and examine the significance and relative magnitude of the estimated parameters (e.g., using p values and measures effect size).

Simulation proceeds similarly except that the model may be in the form of an algorithm or computer program instead of a set of equations, and this model is used to generate simulated data rather than predictions from real data. If possible, the simulated data is compared to available real data to test how similar the two are and assess the validity of the simulation.

A representative example of data simulation in the social sciences is Jager and colleagues' [10] study of group conflict. They used data simulations to study conflict in crowds made up of groups with different allegiances, such as supporters of different football teams. By simulating groups of different sizes and different proportions of aggressive members, they found that conflict was most common when one group was larger than the other and the larger group had a relatively high proportion of aggressive members.

As this example suggests, data simulation has a number of affordances. First, it would be difficult—or at least unethical—to collect data about crowds of people fighting each other. Data simulation allows researchers to generate data that abstracts the situation in a relatively easy and safe way. Second, simulating data provides the researchers with a high degree of control over the design of the experiment. In the example above, Jager and colleagues were able to control the number of data points in each sample and the proportions of aggressive members; they did not have to rely on which participants happened to be available or consent to their study. Relatedly, data simulation allowed them to examine plausible cases that might have gone missed if they had relied on traditional data collection methods—for example, what would happen when the groups of supporters were exactly the same size? In this sense, the simulation allowed them to generalize their findings to a broader variety of situations.

Despite these advantages, at first glance data simulations do not seem to cohere with the QE process, which is so dependent on real data. The dashed links in Fig. 2 indicate steps that are technically unnecessary for data simulation to proceed. While it can be useful to collect real data and use it to assess the validity of the simulation, it is possible (and common) to operate solely on simulated data. However, such an approach is problematic in the context of QE. A QE researcher cannot arrive at qualitative claims when no real qualitative data exists; a researcher cannot close the interpretive loop if there is no real data to return to. In the next sections we overview four examples of applications

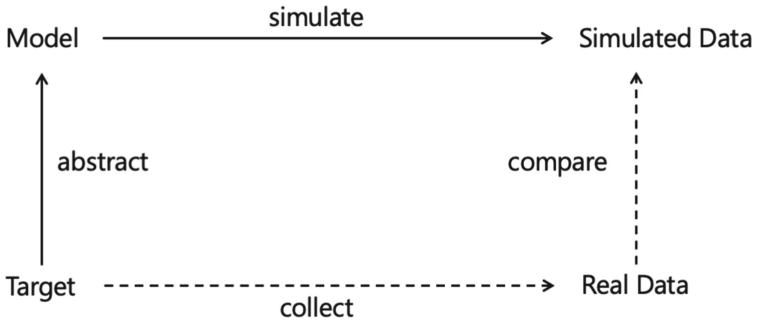

Fig. 2. The data simulation process. Adapted from [9]

of data simulation to QE to argue that despite these differences, data simulation can—and already does—have an important role in QE.

4 Data Simulation in Quantitative Ethnography

4.1 Transcription Error

QE researchers often rely on audio or video recordings of events as data. To analyze these data they typically transcribe it to some other machine-readable format—e.g., text—and then code the transcription [11, 24]. As a result, transcription provides a critical link in connecting events in the world to models and understandings of those events. Unfortunately, transcription processes are imperfect and transcription errors can lead to coding errors, each of which compound to negatively influence the integrity of the subsequent analysis.

To examine the impact of transcription error on coding performance, Eagan [6] used a data simulation. In terms of Fig. 2 above, his target was the relationship between transcription error and coding error—that is, the extent to which errors in a transcription impact the accuracy of labeling data for a [C]ode. To model this target, he investigated three main parameters that had previously been shown to influence coding performance:

- *base rate*: the frequency with which a code appears in a dataset [5, 14, 23].
- *token rate*: the number of unique tokens[1] used to code the dataset divided by the total number of unique tokens in the dataset [1, 4, 12].
- *redundancy rate*: for an individual dataset, the ratio of data segments with multiple independent examples of a code to the total number of positively coded data segments [8].

To examine how these parameters relate to the impact of transcription error on coding performance, Eagan developed the *sensitivity analysis for transcription error* (SATE) method. In statistics, sensitivity analyses measure the level of bias or error that

[1] In textual data, lines of text are composed of tokens: the individual, or unique combinations of, pieces of information that each line of data contains [22].

would need to be present in a dataset to invalidate a given inference, statistical result, or interpretation [7]—that is, the extent to which the data could be altered until an original result becomes invalid. This same approach can be used to examine the impact error has on the inferences or claims made in coding processes.

Eagan used the SATE method to study both real and simulated *data-classifier systems* (DCSs): the pairing of an individual dataset and a specific classifier or coding process. He did so by introducing transcription error to datasets and re-coding them to determine whether the resulting coding error was acceptable. Transcription errors were introduced using a 2-state Markov modulating failure process that goes through a dataset word by word with, in this case, a 5% chance of replacing each word with another word from the dataset.

First, Eagan used the SATE method on 18 DCSs from three different real world learning situations. For each DCS, 5% transcription error was introduced, then the dataset with error was re-coded with the automated classifier and kappa was calculated between the original coding and the coding of the data containing transcription errors. This process of transcription error introduction, re-coding, and kappa calculation was repeated 2,000 times creating a distribution of kappa for each DCS. If 95% of the distribution was greater than a coding performance threshold of kappa equal to 0.9, the DCS was considered robust to 5% transcription error; otherwise it was considered sensitive to 5% transcription error. He used this approach to demonstrated that SATE could discriminate between DCSs that were sensitive to 5% transcription error and those that were robust.

His analyses aligned with the previously specified mechanisms of transcription error influencing coding performance, however the analyses with actual data were too underpowered—that is he did not have enough data to find statistically significant relationships between these mechanisms or their interactions in real data. In addition, while the actual DCSs and prior work provided some guidance as to the ranges of the three parameters of interest, they did not offer examples or representations of all combinations of these parameters researchers could expect to encounter.

To investigate how transcription errors impact coding performance more thoroughly, Eagan created simulated data and associated classifiers to create *simulated DCSs* that are more representative of DCSs researchers could expect to see in the real world. As a result, he was able to assess the significant main effects and three-way interactions between base rate, token rate, and redundancy rate influencing the impact of transcription error on coding performance. In general, as base rate increases sensitivity to transcription error decreases; as token rate increases, sensitivity to transcription error increases; and as redundancy rate increases, some aspects of coding performance increase, but interactions make this relationship more complex (for more details see [6]).

4.2 Shaffer's Rho

The work by Eagan described above used data simulations to examine the relationships among data representations, classifier features, and classifier reliability. This work assumes that there is some defensible way to warrant classifier reliability—that the rate of agreement between two or more raters on some sample of data is suitably high and that the agreement would hold—allowing for some small level of disagreement—if they were to code the rest of the data. As many QE researchers know, this warrant comes in

the form of Shaffer's rho [14]. However, it is likely less well known that the calculation of Shaffer's rho itself relies on data simulation.

To establish the reliability of coding approaches, researchers often use IRR metrics such as Cohen's kappa—especially when there is too much data or not enough time for one or more raters to code all of the data. The basic idea of using IRR metrics is to measure and control the amount of uncertainty—that is, disagreement between raters—involved in a coding process [16]. However, as Eagan and colleagues [4, 5] have argued, the way researchers commonly use IRR metrics is fundamentally flawed. Many researchers compute an IRR metric on a sub-sample of their data and simply assume that it generalizes to the rest of their dataset. That is, they do not control for cases where the IRR in a sample is over a reliability threshold (say Cohen's kappa > 0.65), but the IRR for the entire dataset is below that threshold—a Type I Error.

Shaffer's rho was developed to address this methodological gap. Here, the target of interest is the coding reliability of two raters. Given a real set of coded data, the algorithm that calculates Shaffer's rho simulates two coding processes over some hypothetical dataset where the agreement between the two raters is less than the IRR threshold of interest. In other words, a data simulation is used to generate a large number of coding pairs that are unreliable given some IRR threshold. Critically, this simulated data shares important characteristics with the real data on hand.

Next, a portion of this simulated data is sampled and the IRR measure is calculated. This process of simulating data and calculating IRR on a subsample of the simulated data is repeated hundreds of times and the IRR values from the samples generate a distribution of IRR measurements under a null hypothesis—namely that the observed IRR was sampled from a larger dataset for which the two raters would not have an acceptable level of agreement. If the observed IRR measurement—the measurement obtained by two raters on the real data for the code in question—is greater than 95% of the IRR values in the null hypothesis distribution, a researcher may conclude that their observed agreement generalizes to the rest of their dataset.

4.3 The Expected Value Test

In many QE analyses, the outputs of statistical models are used to as quantitative warrants for qualitative claims about the connections among [C]odes. One common type of model used in QE is ENA, which identifies the co-occurrence of [c]odes within data. [p]arameters derived from ENA can then be tested for statistical significance to warrant theoretical saturation.

As Swiecki [19] argues, these [p]arameters are often derived from the differences between two samples—say patterns of connections pre and post some intervention, or differences in connections between control and treatment groups. However, QE researchers may not always be able—or want—to compare samples. In some cases, they may be interested in the connections among [C]odes in the [D]iscourse of a single sample—say one classroom or one group of students. Swiecki (ibid.) developed a data simulation-based test—the *expected value test* (EVT)—to produce a [p]arameter appropriate for these kinds of single sample cases.

The method relies on comparing an ENA model developed from the real data to a distribution of ENA models developed from simulated data. In typical applications of

ENA, results are derived in terms of the dimensional reduced networks for each unit of analysis (ENA scores); however, for this test, the results are derived in terms of the full networks for each unit. These networks can be thought of as points in a high-dimensional space. Any collection of points has an average called a centroid and two points close together in this space are considered similar—that is the units of analysis corresponding to these points made similar kinds of connections (as identified by ENA). The method includes the following steps:

- Generate an ENA model from the real data on hand (observed model).
- Generate a distribution of ENA models from simulated data in which the codes and order of lines—e.g., turns of talk—have been repeatedly randomized—that is, a distribution of chance-based models.
- Calculate the similarity of the observed model to the average, or centroid, of the chance-based models. Calculate the distribution of similarities of the chance-based models to the centroid. Compare the observed similarity to the distribution.

Here, the data simulation takes the form of randomized data. The logic being that if the connections identified in the real data constituted a systematic pattern in the data, then they should statistically differ from connections identified in randomized sets of that data. Using this method, Swiecki (ibid.) was able to show that the EVT could distinguish between systematic and non-systematic connections in real data, suggesting that the method provided a plausible quantitative warrant for QE analyses of single samples.

4.4 Informational Interdependence

Collaborative problem-solving has been studied extensively by QE researchers (see, for example, [2, 13, 21]). As several researchers argue, collaborative problem-solving is characterized by different types of interdependence among group members. For example, DeChurch and Mesmer-Magnus [3] argue that *informational interdependence* arises when different individuals need to share different kinds of information to complete a task. Swiecki and colleagues [20] developed a data simulation to explore the nature of informational interdependence during collaborative problem-solving.

Prior to designing the simulation they examined data collected from a real world learning situation that made use of a jigsaw pedagogical design [18]. In a jigsaw design, each team is assigned a unique topic on which to become an "expert". After learning about their topic, new teams are formed in which each person has expertise in a different topic. In these new teams, individuals communicate their knowledge of their assigned topics with the others. Because informational interdependence involves the sharing of different information among individuals, the researchers hypothesized that the *interactivity* among teammates and the *dissimilarity* of the information they shared could be used to predict the amount of informational interdependence on the team, and thus, the impact of pedagogical designs like the jigsaw.

The data on hand consisted of digital records of conversations that teams had using an online chat messing tool. These teams were tasked with a mechanical engineering problem, namely, to design an exoskeleton for rescue workers that would perform well

in terms of attributes like cost and user safety. Following a typical QE process, the researchers coded these data for the presence or absence of [C]odes related to engineering design in this context. In particular, they coded for concepts like design inputs and measurable design outcomes. To measure the effect of the jigsaw—and thus the extent of informational interdependence—they also included the jigsaw topics as [C]odes, which represented particular design inputs that individual team members were assigned to learn about before teams were re-formed.

Analyzing the data qualitatively, they found that chats from the pre-jigsaw sample were focused on the relationships between design inputs (other than the jigsaw topics) and design outputs. Chats from the post-jigsaw sample were focused on the relationships among the different jigsaw topics and the design outputs. They also noticed that the post-jigsaw sample was characterized by a higher level of interaction among teammates (individuals tended to exchange turns of talk rather than have monologue-like sequences of chats) and greater focus on sharing different kinds of information. In other words, interactivity and dissimilarity of information seemed to be related to informational interdependence.

To provide a quantitative warrant for these claims, the researchers developed an ENA model of the connections between [c]odes present in the data and compared the connections identified in the pre/post-jigsaw samples. This analysis yielded a statistically significant difference between the two samples that aligned with qualitative findings. A subsequent analysis regressed the ENA scores on the significant dimension on measures of interactivity and dissimilarity, controlling for team membership. The regression model showed that the mean dissimilarity metric of an individual's team was significantly associated with the ENA score, controlling for team effects—individuals on teams that shared different kinds of information tended to talk more like post-jigsaw teams. Put another way, sharing different kinds of information was positively related to informational interdependence.

While these results were useful, they were limited by the nature of the data on hand. Combinations of dissimilarity and interactivity were only present for limited ranges, raising the question of what the relationship among informational interdependence, dissimilarity, and interactivity would be if more complete data was available—that is, combinations throughout the range of both variables. To investigate this question, the researchers developed a data simulation.

The original data was the actual chat messages sent by the participants. These messages were coded for particular categories and then the relationships between these codes were modeled. Because it would be too difficult (and nonsensical) to attempt to simulate chat messages themselves, the researchers simulated the patterns of codes present in messages instead. Doing so required two generating mechanisms: one that determined the order in which the simulated participants "chatted" and the other to determine the codes present in their "chats".

To generate the sequence in which the simulated participants chatted, the researchers used a lag-1 transition matrix. To generate the codes present in the chats, they used a co-occurrence probability matrix for each simulated participant that was based on the their observed data. Given a pair of participants and a prior chat, the probabilities in

these matrices determined which codes would be present or absent in the subsequent chat.

After validating the simulation by comparing its output to the actual data (see [20] for details), the researchers were able to simulate data under a larger variety of dissimilarity and interactivity combinations than were present in the real data and test the effect of these metrics on informational interdependence, which, as with the real data, was operationalized in terms of their location on the significant dimension of the original ENA space. The results suggested that dissimilarity and interactivity at the team *and* individual levels were significantly related to informational interdependence, expanding the results of the analysis of the real data.

5 Discussion

Thus far we have argued that QE analysis are characterized by two important features. First, QE analysis are simplifications of observed phenomena and are thus prone to error. Second, QE is an analytical process that fundamentally relies on the relationships among qualitative claims, quantitative representations, and real data. We have also given a brief overview of the application of data simulations in the context of QE. What remains is to explicitly link these examples of data simulation to the QE framework. These links are summarized in Fig. 3 and expanded upon below.

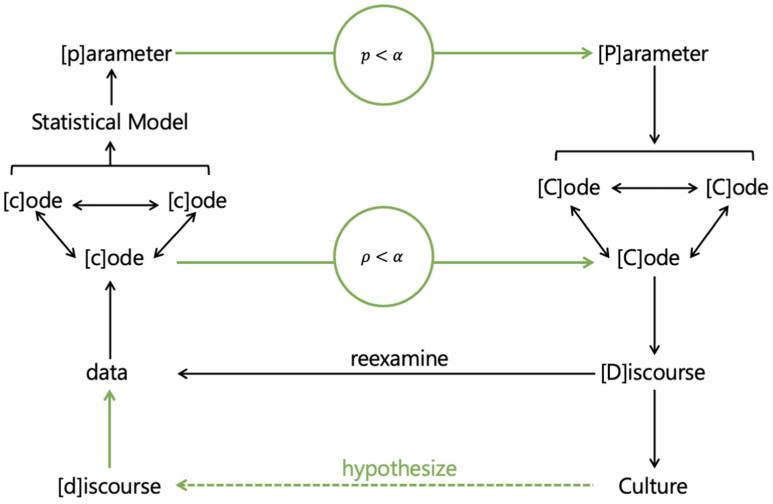

Fig. 3. The QE process with links to data simulation in green.

5.1 Link 1: [D]iscourse to Data

QE researchers record a subset of the [d]iscourse they observe as data. A common practice is to make audio or video recordings of [d]iscourse and then transcribe those

recordings for analysis. Any transcription process, whether manual or automated, is prone to error. Words are mistaken, missed, or attributed to the wrong speaker. Because QE researchers code these data to make claims about the [D]iscourse of some culture, these errors, if left uncontrolled, can damage the coding, the analysis of connections, and consequently, the validity of the entire analysis. Put simply, if the data are bad to begin with, the analysis will be bad as well. Garbage in, garbage out.

The SATE method provides a means to investigate the tolerance of the coding process to transcription errors and guidance on how to control for transcription errors. By simulating a range of coding schemes and error prone data, the method can identify specific features that QE researchers can use to examine whether transcription errors in their data are likely to negatively impact the reliability of coding schemes applied to those data. In particular, [c]odes that are prevalent in a dataset and tend to appear multiple times within segments of data tend to be more robust to transcription error; [c]odes that are relatively unique compared to the data tend to be more sensitive to transcription error.

5.2 Link 2: [C]odes to [C]odes

QE researchers—as well as ethnographers more generally—come to understand the [D]iscourse of some culture by understanding the [C]odes of that culture—the kinds of things they say, do, feel, think, and believe that define them. To do so, they identify [c]odes in their data—actual pieces of data that they use to argue for the presence (or absence) of [C]odes. This act of pointing is reductive in the sense that it takes a high-level, sometimes vague and nuanced concept, and materializes it in the form of pieces of data. As a reductive act, it is prone to error and open for disagreement between parties—multiple researchers say, or the researcher and members of the culture.

To justify the link between [c]odes and [C]odes, QE researchers seek some measurable consensus or agreement. In many cases they code a subset of the data, measure the agreement, and determine if the agreement is good enough for the purposes of the analysis. Shaffer's rho—which is derived, in part, from a data simulation—provides evidence that the agreement reached on this sample of data would generalize to the rest of the data at hand, and thus supports and strengthens the link between [c]odes and [C]odes.

5.3 Link 3: Connections

[C]odes in isolation do not define the [D]iscourse of a culture; it is the relationships among [C]odes that allows us to understand the culture in some way. In turn, QE researchers often seek to identify the relationships among [c]odes in their data and warrant that these relationships are indicative of the patterns among [C]odes that help to define the culture. One way to warrant this link is to use ENA to identify co-occurrences between [c]odes in the data and then perform a statistical test, the result of which can suggest that the patterns observed in the data are systematic—signal, not noise.

While there are established ways of running these statistical tests for cases in which that are two samples of data that researchers wish to compare, the way forward is less clear when researchers can or want to only describe the nature of a single sample. Using a simulation to compare the connections observed in the data to the connections identified in a randomized version of that data, the EVT provides a statistical test of the connections

between [c]odes for single samples. In turn, this method can support the link between the connections among [c]odes and the connections among [C]odes that characterize the [D]iscourse of some culture.

5.4 Link 4: Building Theory

Initiating a QE analysis requires observations of the [d]iscourse of a culture. These observations are recorded as data, and the analysis proceeds. When the analysis is finished, if we have done it well, we have evidence that the qualitative claims we are making are systematic properties of the [D]iscourse of the culture we are studying—we have evidence of theoretical saturation. These claims, however, are inherently limited by the [d]iscourse we have observed.

Although QE uses statistical tests, their function is not the same as in typical quantitative inquiry. Outside of QE, statistical tests are typically used to generalize claims made about some sample of observations—on people, say—to the larger population of observations we could have made about other people. In other words, typical statistical tests warrant generalizations *outside* the data we have. In QE, statistical tests are used to generalize claims made about some sample of observations on people to the larger population of observations we could have made about the same people. In QE, statistical tests warrant generalizations *within* the data we have.

Nonetheless, the data we have may be severely limited and thus our claims narrow. Data simulations can help to expand upon the data we have. As shown by the work of Swiecki and colleagues [20] as well as Eagan [6], data simulation can produce results that may have been missed if only real data had been examined. However, these results do not necessarily expand the kinds of claims we can make in the context of QE. The reason being that they are initially unverifiable.

A QE analysis is not finished when a significant statistical result is or is not obtained. The analysis is finished when the researchers re-examine the claims that they have supported or refuted in terms of the [d]iscourse they have recorded as data—that is, when they have closed the interpretive loop. When data in question is simulated, closing the loop is not possible—there is no qualitative account to check the results against. Of course, this does not mean that results from simulated data are useless. Instead, it reframes these results as the starting point for subsequent analysis; the results become hypotheses that can be tested by observing more [d]iscourse and conducting future QE analyses. In other words, they become mechanisms for testing and building theories about some culture.

6 Conclusion

In this paper, we provided an overview of the QE process and examples of the use of data simulation in QE. We argued that QE is reliant on real—that is, actually observed—data and that QE is error prone. In spite of the former and because of the later, we argued that data simulation has a role to play in QE. A look at Fig. 3 suggests that this role is more than just a cursory one. New links in the QE process can be created and existing links can be reinforced.

Our work here is limited in the sense that we have only provided evidence for the relationships between QE and data simulation. We have not discussed the major practical issues associated with using data simulations in QE nor have we provided guidelines for implementing simulations in the QE context. This paper is a prerequisite to that future work. For now, we hope that we have given insight to the usefulness of data simulation in QE and that this paper will spark debate and study about whether and how data simulation should be incorporated into future QE work.

Acknowledgements. This work was funded in part by Monash University, the National Science Foundation (DRL-1661036, DRL-1713110, DRL-2100320), the Wisconsin Alumni Research Foundation, and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.

References

1. Almuallim, H., Dietterich, T.G.: Learning with many irrelevant features. In: Proceedings of the Ninth National Conference on Artificial Intelligence AAAI, vol. 91, pp. 547–552 (July 1991)
2. Csanadi, A., Eagan, B., Kollar, I., Shaffer, D.W., Fischer, F.: When coding-and counting is not enough: using epistemic network analysis (ENA) to analyze verbal data in CSCL research. Int. J. Comput.-Support. Collab. Learn. **13**(4), 419–438 (2018)
3. DeChurch, L.A., Mesmer-Magnus, J.R.: The cognitive underpinnings of effective teamwork: a meta-analysis. J. Appl. Psychol. **95**(1), 32–53 (2010). http://doi.org/10.1037/a0017328, http://doi.apa.org/getdoi.cfm?doi=10.1037/a0017328
4. Eagan, B., Brohinsky, J., Wang, J., Shaffer, D.W.: Testing the reliability of inter-rater reliability. In: Proceedings of the 10th International Conference on Learning Analytics & Knowledge, pp. 454–461. Association for Computing Machinery (2020). http://www.epistemicanalytics.org/wpcontent/uploads/2020/06/LAK20_Eagan_IRR_Camera_Ready.pdf,https://doi.org/10.1145/3375462.3375508
5. Eagan, B., Rogers, B., Serlin, R., Ruis, A., Arastoopour, G., Shaffer, D.W.: Can we rely on reliability? Testing the assumptions of inter-rater reliability. In: Smith, B., Borge, M., Mercier, E., Yon Lim, K. (eds.) Making a Difference: Prioritizing Equity and Access in CSCL: 12th International Conference on Computer Supported Collaborative Learning (CSCL) 2017, vol. 2, pp. 529–532 (2017)
6. Eagan, B.R.: Measuring the Impact of Transcription Error. Doctoral Dissertation, University of Wisconsin - Madison (2020)
7. Frank, K., Min, K.S.: 10. Indices of robustness for sample representation. Sociol. Methodol. **37**(1), 349–392 (2007)
8. Gilad-Bachrach, R., Navot, A., Tishby, N.: Margin based feature selection-theory and algorithms. In: Proceedings of the Twenty-First International Conference on Machine learning, p. 43 (July 2004)
9. Gilber, N., Troitzsch, K.: Simulation for the Social Scientist. McGraw-Hill Education, UK (2005)
10. Jager, W., Popping, R., Van de Sande, H., Jager, W., Popping, R., Van de Sande, H.: Clustering and fighting in two-party crowds: simulating the approach avoidance conflict. J. Artif. Soc. Soc. Simul. **4**(3), 1–18 (2001)

11. Kaliisa, R., Misiejuk, K., Arastoopour, G., Misfeldt, M.: Scoping the emerging field of quantitative ethnography: opportunities, challenges and future directions. In: Ruis, A., Lee, S. (eds.) Advances in Quantitative Ethnography: Second International Conference, ICQE 2020, Malibu, CA, USA, February 1–3, 2021, Proceedings, pp. 3–17. Springer, Heidelberg (2021). https://doi.org/10.1007/9783-030-67788-6_1, https://link.springer.com/chapter/10.1007/978-3-030-67788-6_1
12. Liu, H., Setiono, R.: A probabilistic approach to feature selection-a filter solution. In: ICML, vol. 96, pp. 319–327 (July 1996)
13. Ruis, A.R., Siebert-Evenstone, A.L., Pozen, R., Eagan, B.R., Shaffer, D.W.: Finding common ground: a method for measuring recent temporal context in analyses of complex, collaborative thinking. In: A Wide Lens: Combining Embodied, Enactive, Extended, and Embedded Learning in Collaborative Settings: 13th International Conference on Computer Supported Collaborative Learning (CSCL), vol. 1, pp. 136–143 (2019)
14. Shaffer, D.W.: Quantitative Ethnography. Cathcart Press (2017). http://www.quantitativeethnography.org/
15. Shaffer, D.W., Collier, W., Ruis, A.R.: A tutorial on epistemic network analysis: analyzing the structure of connections in cognitive, social, and interaction data. J. Learn. Anal. **3**(3), 9–45 (2016). http://learninganalytics.info/journals/index.php/JLA/article/view/4329
16. Shaffer, D.W., Ruis, A.R.: How we code. In: Ruis, A.R., Lee, S.B. (eds.) ICQE 2021. CCIS, vol. 1312, pp. 62–77. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67788-6_5
17. Shaffer, D.W., Serlin, R.: What good are statistics that don't generalize? Educ. Res. **33**(9), 14–25 (2004)
18. Slavin, R.E.: Cooperative Learning. Learning and Cognition in Education, pp. 160–166. Elsevier Academic Press, Boston (2011)
19. Swiecki, Z.: The expected value test: a new statistical warrant for theoretical saturation. In: Wasson, B., Zörgő, S. (eds.) Advances in Quantitative Ethnography: Third International Conference, ICQE 2021 Virtual Event, November 6–11, 2021, Proceedings. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-030-93859-8_4
20. Swiecki, Z., Marquart, C., Eagan, B.: Simulating collaborative discourse. In: Paper Accepted to the ISLS Annual Meeting 2022 (2022)
21. Swiecki, Z., Ruis, A.R., Farrell, C., Shaffer, D.W.: Assessing individual contributions to collaborative problem solving: a network analysis approach. Comput. Hum. Behav. **104**, 105876 (2020). https://doi.org/10.1016/j.chb.2019.01.009
22. Webster, J., Chunuy, K.: Tokenizaion as the initial phase in NLP. In: COLING 1992 Volume 4; the 14th International Conference on Computational Linguistics (1992)
23. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, USA (2011)
24. Zörgő, S., Peters, G.J.Y., Porter, C., Moraes, M., Donegan, S., Eagan, B.: Methodology in the mirror: a living, systematic review of works in quantitative ethnography. In: Wasson, B., Zörgő, S. (eds.) Advances in Quantitative Ethnography: Third International Conference, ICQE 2021 Virtual Event, November 6–11, 2021, Proceedings. Springer, Heidelberg (November 2021). https://doi.org/10.1007/978-3-030-93859-8_10