# Lost In Translation:
# How Not to Make Qualitative Research More Scientific

Mario Luis Small
Princeton University
*National Science Foundation Workshop on Qualitative Methods*

**Recommended citation**: Small, Mario L. Forthcoming. "Lost in Translation: How Not to Make Qualitative Research More Scientific." In Michèle Lamont and Patricia White (editors), *Report from Workshop on Interdisciplinary Standards for Systematic Qualitative Research*. Washington, DC: National Science Foundation

## INTRODUCTION

One version of an old joke among Spanish-speaking immigrants tells of Pablo running into his friend María, who is with a group of friends. Everyone speaks Spanish. Pablo, who has been in the U.S. for 10 years, says, "María, introdúceme a tus amigos." The joke is that the phrase does not mean "introduce me to your friends"; it means something closer to "insert your friends inside me." The right phrase is "preséntame a tus amigos," but Pablo, accustomed to the English language, has gotten his languages mixed up. The joke is funny because everyone knows what Pablo meant to say. In other settings, however, the problems of translation can lead to much worse than embarrassing misunderstandings.

Methods of scientific inquiry are languages to the extent that they constitute systems of thought, with terms that have specific meanings and ways of framing problems that make sense only within the system. Most quantitative researchers employ the language of frequentist or classical statistics; qualitative researchers often employ the language of participant observation or the different language of in-depth interviewing. As Howard Becker insists, all methods fundamentally seek the same things (that arguments are backed by data, that procedures can in theory be repeated, etc.); however, the languages with which they say those things are significantly different, and some statements only make sense in some languages.[1]

If methods are languages, then the most important issue facing qualitative researchers---especially those concerned about the science of their work---is translation. This is especially a problem for researchers who study topics, such as neighborhood poverty, where both quantitative and qualitative research is necessary. Many social scientists in recent years have rightly attempted to bridge the gaps between qualitative and quantitative thinking. But many of these attempts have involved making qualitative research, for which there are fewer agreed-upon rules, come closer to matching the logic of inquiry of classical statistics. Thus, qualitative researchers encourage their students to make sure their samples are "representative" or "unbiased," and

[1] In this light, it is worth noting that even quantitative researchers have different languages. Classical or frequentist statistics has a common set of terms (central limit theorem, null hypothesis, t-statistic) tied to a way of framing problems different from that of Bayesian statistics, whose own set of terms (prior distribution, posterior distribution, non-informative prior) is tied to its way of framing problems.

quantitative ones warn against "selecting on the dependent variable." I believe that in doing this we are forcing words into systems of thought where they do not belong, and that this will exacerbate, rather than improve, our problems of communication.

I hope to make this problem clear by focusing on two examples of how ethnographers concerned about science attempt to make their work "generalizable." My own work is in the fields of inequality and urban poverty, and the translation problems I discuss refer to those between qualitative and quantitative researchers in inequality, simply because this is the work I know best. Some of my work is qualitative and some of it is quantitative; I write as an interpreter hoping to increase the quality of translation, not as a chastiser of one method or another, anymore than one would chastise Spanish for not being English.

FIRST EXAMPLE: IN-DEPTH INTERVIEW

Jane is writing her second-year paper on the attitudes of working-class African-Americans about immigration. She wants to conduct in-depth interviews to capture the nuances of these attitudes, and is planning to interview 35 respondents. Jane worries, "Will my findings be generalizable to the wider population?"

Her adviser, a qualitative researcher, recommends finding a city with a large working class African-American population, obtaining a telephone directory, and randomly sampling people from it. He knows to expect, at best a 50% response rate, so he recommends contacting 100 people selected at random. Jane follows the plan, and miraculously all phone numbers are valid and everyone answers the phone. Of the 100 respondents, 60 hang up on her, 40 agree to an interview, and 35 follow through with it. (Interviewers recognize that, for some populations, these are wildly optimistic figures.) She conducts 35 high-quality 2-hour interviews that delve deeply into attitudes about immigration, uncovering subtle causal relationships among attitudes, experience with discrimination, gender, and Southern origins, and she happily completes her paper. Since her method mirrors that of Lamont's (1992) well regarded *Money, Manners, and Morals*, she is confident in the "generalizability" of her findings about the black working class.

The problem is that under no statistical definition of generalizability can the responses of those 35 individuals be considered to reflect reliably the conditions of the African-American working class. In fact, a quantitative researcher's confidence in Jane's estimates would be just about the same had she simply gone to any neighborhood in the city, and interviewed the heads of households of the first 35 houses on Main Street. It is tempting to think that the first sample is significantly better because it is "more random" but it hardly a statistical improvement.

There are two reasons for this. First, the sample has an inbuilt and unaccounted for *bias*.[2] Jane only interviewed the 35% of respondents who were polite enough to talk to her, friendly enough to make an appointment based on a cold-call from a stranger, and extroverted enough to share their feelings with this stranger for 2 hours. It is very likely that these people have systematically different attitudes about others, including immigrants, than non-respondents. Since we do not

[2] For a discussion of these issues from researchers aimed at bridging the qualitative/quantitative divide, see King, Keohane, and Verba (1994:63ff).

know anything about those who did not respond (they hung up), we have no way of adjusting the inferences we obtained from the 35 respondents. In addition, since we do not know anything about working class blacks in other cities, we do not know if, had she had 100% response rates, respondents in her city were typical or atypical of the black working class.

Second, regardless of how it was selected, the sample is *too small* to make confident predictions about complex relationships in the population of working-class blacks at large. Many students ask, "How many people do I need to interview for my findings to be generalizable?" The answer depends on the distribution of the variables of interest, whether the students want to describe distributions (e.g., proportion Democrat) or present causal relationships (e.g., whether Republicans will have stronger anti-immigrant attitudes than Democrats), and how many variables are involved, among other things. King, Keohane, and Verba (1994:213) provide a formula, based on standard statistical assumptions. The short answer, however, is that rarely will students have enough well-selected in-depth interview respondents that their findings about subtle causal relationships involving multiple variables will be statistically generalizable to a large national population. For that, one needs a survey.

Suppose Jane only wanted to know how many working class blacks are pro-immigration reform (one Y/N question); and she wanted to be 95% confident that the average in her sample matched the average in the population at large within +/- 5 percentage points; and that the population of working-class blacks in the U.S. were 2,000,000 people (for large populations, the exact size does not matter very much). In this case, she would need 385 respondents.[3] If Jane narrowed her focus, and only wanted to be confident about the 1,000 working-class blacks in one city, she would need 278.

Some qualitative researchers prefer to ignore these issues and refer to studies such as Jane's as "representative," but the truth is that in doing so qualitative research is simply adopting words without adopting their meaning.

The natural question is whether, having acknowledged this, it is still not better for Jane to have picked her respondents "at random" (in quotation marks because her final sample is not statistically random due to high non-response) than engaging in some other non-random selection strategy. Not always. Consider, for example, sampling for range (Weiss 1994). Suppose Jane suspected strongly that gay and lesbian respondents would be more sympathetic to immigrants. Even a truly random sample would have yielded, at best, 3 or 4 gay or lesbian respondents out of 35, of which 1 or 2, at best, would reveal this to her. This would leave her no room to examine this question. In these circumstances, Jane would be better served designing her study to include a large, pre-determined number of gay and lesbian respondents, even if this meant finding them through non-random means, such as organizations. For many questions of interest to interview-based sociologists, sampling for range is a more effective strategy.

---

[3] The formula is $n = (Z^2 * p * (1-p))/C^2$, where Z is the Z value (1.96 for a 95% confidence level); C is the confidence interval, expressed as a decimal (in this case .05); and p is the percentage of people who are expected to be, in this case, pro-reform. We assume .5, the most conservative assumption. If 51% are pro- and 49% are anti-reform, the room for error is high, so a large sample is needed; if 90% were pro-reform one could get by on a much smaller sample of 139. There are dozens of sample size calculators on the web, where one can manipulate the assumptions. For example, www.raosoft.com/samplesize.html.

Even in circumstances where researchers are not seeking a particular and small population, random is not necessarily better. Snow-ball sampling, for example, involves asking respondents to recommend other respondents. This has the well-known problem that respondents will tend to be in-network members. Because of this "bias," some researchers are reluctant to recommend this method over random sampling. But snow-balling almost always leads to higher response rates, since people are less reluctant to speak to strangers when they are sent from a trusted source. So, which is worse---the "bias" from in-network sampling or the "bias" from low response rates in random selection? Neither is; which method to employ depends only on the objectives of the project.

SECOND EXAMPLE: NEIGHBORHOOD STUDY

There is a similar problem in participant observation research aimed at dealing with large-n questions. Bill wants to study how neighborhood poverty affects out of wedlock births, by conducting an in-depth ethnography of a single high-poverty neighborhood. His main concern is the set of mechanisms underlying this process, but he wants to make sure his findings are generalizable to all poor neighborhoods. Thus, he does what King, Keohane, and Verba (1994:67-68) might do:

> For example, we could first select our community very carefully in order to make sure that it is especially representative of the rest of the country…. We might ask a few residents or look at newspaper reports to see whether it was an average community or whether some nonsystematic factor had caused the observation to be atypical…. This would be the most difficult part of the case-study estimator, and we would need to be very careful that bias does not creep in. Once we are reasonably confident that bias is minimized, we could focus on increasing efficiency. To do this, we might spend many weeks in the community conducting numerous separate studies…[4]

Bill looks to the census, finds a neighborhood that is 40% poor, 60% black, with 80% of the households female headed and (he discovers at the site) most streets littered and covered in graffiti, all of which seem to accord with his definition of a "representative" poor neighborhood. Bill conducts his study, and finds that the high level of poverty has made residents distrustful of each other. This distrust, he finds, makes the women unwilling to marry the fathers of their children. Since his neighborhood was representative, Bill is confident that neighborhood poverty increases out of wedlock births in poor neighborhoods at large through the mechanism of lower trust.

The problem with this way of thinking is that, no matter how Bill selected his single neighborhood it will never be truly representative of poor neighborhoods. The neighborhood's conditions may happen to match the traits that, from the census, one knows to be at the mean. But, as Frankfort-Nachmias and Nachmias (2000:167) write, "a sample is considered

[4] In this passage, the authors were discussing much broader issues, so this selection does not do justice to their book. The purpose here is not to produce a full-fledged critique of the authors' (in many ways excellent) book. Rather, it is to show the pitfalls of this particular way of thinking about case study selection, which the authors do share with many others.

representative if the analyses made using the sampling units produce results similar to those that would be obtained had the entire population been analyzed." No "sample" of a single neighborhood can match this criterion.

Even obtaining copious and very detailed information on the neighborhood (a generally sensible recommendation by King, Keohane, and Verba [1994]) does not change this fact. Suppose that instead of neighborhoods we were speaking of individuals, and we selected one person with the characteristics of the average American: a married 37-year old female with a high school education who earned $35,038 last year.[5] We interviewed this female for 2 hours about her opinions on the admission of Turkey into the European Union. How confident would we be that her thoughts accurately reflected those of the average American? A scientist would have no confidence, and interviewing her for 20 or 200 additional hours would not change this.

Bill does not have a "sample" of 1; he has a single case study. Suppose that Bill had chosen a neighborhood with a 40% poverty rate but with no garbage or graffiti and a unique architectural design due to the influence of a mayor interested in promoting architecture in the city. It is tempting to think that inferences based on the second case would be less statistically generalizable but, based on a sample of 1, they are neither more nor less so. As before, one could ask if there is any harm in going for the statistics-inspired "random" or "average" strategy. Sometimes there is. Suppose the mayor in the second case also had a radical and unique policy whereby mothers received significantly higher rent subsidies plus $1,000 per child for a college fund if they married before the birth of their second child. This rare case would suddenly present Bill an exceptional opportunity to examine the relationship among high poverty, policy, and out of wedlock births in ways that cases that happen to be at the mean might not.[6] In case studies, rare cases are often precisely what the researcher wants (Yin 2002). In all case studies, though, selection must be made primarily on substance.

## GIVE UP?

My purpose here is not to argue that ethnographic or interview-based methods are destined to be unscientific. On the contrary, I strongly believe that in-depth interviewing and participant observation constitute two of the purest *empirical* methods there are, which is why I rely on them in my work. (Statistical surveys rely on abstractions of the world into pre-determined variables, and thus are inherently once-removed from empirical reality.) My objective is to encourage qualitative researchers to produce scientific work based on their own language, not that of others.

Consider Lamont's (1992) aforementioned book, a study of 160 upper-middle class men in France and the United States. I think Lamont's book is one of the most methodologically sophisticated interview-based studies in recent years. However, I do not believe, as others have commented, that it is sophisticated because "she had a representative sample." The study's

---

[5] The median age for males and females is 37; more individuals are married than never married, widowed, or divorced; among persons 25 or older, more are high school graduates or graduates with some college than not high school graduates, college graduates, or persons with advanced degrees; $35,038 is the median earnings for individuals for the year 2002. See Section 1, Population, of the *Statistical Abstract of the United States*, http://www.census.gov/prod/2004pubs/04statab/pop.pdf .

[6] These arguments are discussed further in the concluding chapter of my (2004) *Villa Victoria: The Transformation of Social Capital in a Boston Barrio*.

response rate was very low, between 42% and 58%, by liberal estimates[7] (Lamont 1992: 218). In addition, the samples are small, only 80 individuals in each country (40 in each site). This is common among field-base studies, and not a problem given the arguments the book makes. But pretending it is truly representative only detracts from the true strengths of the work, and encourages young scholars (like Jane) to focus on making qualitative research more quantitative instead of on improving the way in which they handle qualitative research. The methodological sophistication of the book comes from the sensitivity of the interview process; Lamont's ability to interpret the meaning of respondents' statements within their cultural contexts; her use of a comparative model to sharpen her concepts; her judicious use of both semi-structured interviews, which allow findings to emerge inductively, and a structured survey, which provides comparative data across the cases; her thoughtful selection of research sites (Lamont 1992: Appendix II); and her effective use of these data to tell a compelling story.[8]

## ONE ALTERNATIVE: CASES, NOT SAMPLES

Behind the desperate search for "representative" qualitative data in Bill and Jane's projects is the assumption that if one cannot make statistical statements about the distribution of a variable, one is not engaging in science. I believe this is false. Consider psychological experiments. Most of these are conducted on small and highly unrepresentative samples of college students at large research universities. Yet a recent Nobel was awarded for precisely this type of work.

One way to think about alternative conceptions of scientifically rigorous qualitative work is adopting Yin's (2002) distinction between *case study* logic and *sampling* logic. Yin's work is on case studies, but I believe it is applicable to in-depth interview-based studies, which can be seen, rather than as small-*sample* studies, as multiple-*case* studies. In what follows, I am extrapolating from his work. I cannot do justice to his work in these few pages, but one example should suffice.

Sampling and case study logic approaches are different and fully independent ways of approaching data. In a sampling model, the number of cases is predetermined; the sample is meant to be representative; all individuals should have equal (or known) probability of selection; and all units should be subject to exactly the same questionnaire. In a case model, the number of cases is unknown until the study is completed; the collection of cases is, by design, not representative; each individual has its own probability of selection; and different people have different questionnaires. Case study logic is critical when asking *how* and *why* questions, with which a sampling logic has greater difficulty.

An example from a different method (experiments) will show the fruitfulness of a case study approach. Alfonse conducts an experiment in which one group of black and white students at Berkeley is told they will receive an IQ test and another is told nothing. Both complete the same

---

[7] As Lamont writes, the figures "do not include potential respondents who did not provide the information necessary to determine whether they qualified or not" for the study (Lamont 1992:285). Thus, the figures could overstate the response rate.

[8] To be clear, I do not think she was *mistaken* in employing a random sampling strategy. The point is that, if we were to judge it by the (inappropriate) standards of statistical generalizability, the sample is no better than many other alternatives, neither of which would fare very well. One cannot expect high response rates when conducting in-depth interviews regarding personal issues for 2 hours.

test, and blacks in the first group do much worse than whites, while those in the second do as well as whites in their group. Alfonse concludes the fear of fulfilling a stereotype about low IQs among blacks is at play.[9] He then does two things, literal and theoretical replication (Yin 2002). With a colleague at Duke, he repeats the experiment among Duke undergraduates (literal replication); back at Berkeley, he repeats it, but using men and women instead of blacks and whites (theoretical replication). If the theory is right, it should work for anyone, not just blacks and whites. Some results confirm his findings; others do not. Then he tries it among Asians and whites, and among issues other than IQ, and on more campuses, and with high school students and seniors, and on an on. Slowly, as the number of experiments increases, his confidence that his theory is right begins to increase. Eventually, every new experiment contributes very little new knowledge, such that the 89th experiment, with immigrant Asians and Russians in a low-income high school, shows exactly what he expected. At this point, he has attained saturation, and he stops.

Alfonse has just conducted (after many years) a type of multiple-case study. Notice that at no point did Alfonse conduct a random sample of any kind. On the contrary, the characteristics of respondents in every experiment were deliberately chosen.

I suggest that this approach may be used to think about in-depth interview studies. The key is to think about every individual as a single experiment. Jane, without knowing how many respondents she will eventually interview, interviews one. The person recounted experiencing discrimination from Latino immigrants when she was a child, thus developing anti-immigrant sentiments and favoring immigration reform. From the interview, Jane theorizes that blacks who have been discriminated against by Latino immigrants will favor immigration reform. She then searches for blacks who report discrimination from Latinos (literal replication), as well as those who have not experienced it (theoretical replication) and those who experienced discrimination from Russian immigrants (theoretical replication). Importantly, she alters each new interview to make sure to include increasingly refined questions about different aspects of discrimination. She does this over and over. Her last interviews are longer than the first, and they include many more subtle variations on the way one experiences discrimination. Eventually, each new interview is telling her very little she had not already heard about the relationship between discrimination and immigrant attitudes. She has attained saturation.[10]

Jane's method violated nearly all of the tenets of (frequentist) sampling logic. Her group of respondents is not representative; each respondent received a slightly different questionnaire; there was no attempt to minimize statistical bias. Thus, Jane can make no statement about the distribution of attitudes. She cannot report accurately that 25% of working class blacks favor immigration reform, just as Alfonse would not report that 80% of blacks are susceptible to stereotype threat, or that, since 75% of the experiments confirmed his theory, his theory is right 75% of the time (this would be wrong on many, many counts). However, we would have the same confidence in her findings as we do in Alfonse's statements that stereotype threat reduces

[9] This example is (very) loosely based on the work of Claude Steele and colleagues, at Stanford University.

[10] These descriptions of the research process are stylized, as all of them are forced to be. In real life, Jane would have interviewed 10 people before any semblance of a story might have emerged. She then would have limited the scope of her study and her questions, to prevent continuing to interview indefinitely.

performance. Jane's work would be as scientific as Alfonse's, even though neither of them can make distributional statements.[11]

## CONCLUSION

Adopting a case-based logic is *not* the only way to conduct ethnographic research that is scientifically rigorous. The point is that it is possible to conduct rigorous research without employing the assumptions of classical statistics in any way. (In fact, the method described above, in its reliance on revising assumptions based on new information conducted during the study, bears some parallels to Bayesian statistics.) To claim that studies such as Bill's are "generalizable" is to adopt terms while ignoring their meanings. It is to mistake "insert your friends inside me" for effective communication. The strengths of qualitative work come from understanding *how* and *why*, not understanding *how many*, and improving this work should mean improving the reliability of its answers to how and why questions. For qualitative researchers to attempt to make their work statistically representative is to engage in a losing race, one in which those who have large samples, by design, will always win. It is the equivalent of evaluating success in one language on the basis of the grammar and vocabulary of another. In science, many tongues are better than one.

## REFERENCES

Frankfort-Nachmias, Chava and David Nachmias. 2000. *Research Methods in the Social Sciences*. 6th edition. New York: Worth Publishers.

King, Gary, Robert O. Keohane, Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. New Jersey: Princeton University Press.

Lamont, Michèle. 1992. *Money, Morals, & Manners: The Culture of the French and American Upper-Middle Class*. Chicago: University of Chicago Press.

Small, Mario Luis. 2004. *Villa Victoria: The Transformation of Social Capital in a Boston Barrio*. Chicago: University of Chicago Press.

Weiss, Robert. 1994. *Learning from Strangers: The Art and Method of Qualitative Interview Studies*. New York: Free Press.

Yin, Robert. 2002. *Case Study Research: Design and Methods*. Thousand Oaks, CA: Sage.

[11] Of course, if Jane had simply selected her 35 cases as she had before, she would not be able to make statements about distributions either.

Quantitative research is more preferred over qualitative research because it is more scientific, objective, fast, focused and acceptable. However, qualitative research is used when the researcher has no idea what to expect. It is used to define the problem or develop and approach to the problem. Return to Article Details Translation in Qualitative Social Research: The Possible Impossible. Volume 14 , No. 2 , Art. 13 â€“ May 2013.Â  Abstract : In an increasingly globalized world of research, communicating with scholars in the same language and culture and with scholars from other cultures and linguistic background is a sine qua non in/of all sciences, including those using qualitative social research.Â  This is made explicit when current speaker formulates what is happening by using the phrase "in other words," and then uses a different set of words that we are to hear the same as what has been said before. To return to the present situation, why would a teacher rephrase a student's contribution to classroom talk? Many students ask, â€œHow many people do I need to interview for my findings to be generalizable?â€ The answer depends on the distribution of the variables of interest, whether the students want to describe distributions (e.g., proportion Democrat) or present causal relationships (e.g., whether Republicans will have stronger anti-immigrant attitudes than Democrats), and how many variables are involved, among other things.Â  The purpose here is not to produce a full-fledged critique of the authorsâ€™ (in many ways excellent) book. Rather, it is to show the pitfalls of this particular way of thinking about case study selection, which the authors do share with many others.Â  My objective is to encourage qualitative researchers to produce scientific work based on their own language, not that of others. 2005. â€œLost in Translation: How Not to Make Qualitative Research More Scientific.â€ Princeton University. Whyte, William Foote.Â  Pick one puzzle that most interests you and provide a reason for why a qualitative methods approach should, or could productively, be used to find an answer to this puzzle. Week 7 (November 9th) Archives Vitalis, Robert. â€œArchives.â€ The article considers the issues related to the semantic, grammatical, stylistic and technical difficulties currently present in machine translation and compares its four main approaches: Rule-based ( RBMT ), Corpora-based ( CBMT ), Neural ( NMT ), and Hybrid ( HMT ). It also examines some â€œopen systemsâ€, which allow the correction or augmentation of content by the users themselves (â€œcrowdsourced translationâ€). The authors of the article, native speakers presenting different countries (Russia, Greece, Malaysia, Japan and Serbia), tested the translation quality of the most representative phrases from the English, Russian, Greek, Malay and Japanese languages by using different machine translation systems: PROMT (RBMT), Yandex.Translate (HMT) and Google Translate (NMT).