Qualitative Research in the Positivist-Behavioral Tradition

Resources for Addressing Type I and Type II Errors in Code Associations Using ATLAS.ti

Corey M. Abramson

*Corey M. Abramson (***http://cmabramson.com***) is a PhD candidate in the* **Department of Sociology** *at the University of California, Berkeley and a Research Analyst at the University of California San Francisco's Institute for Health Policy Research. His research uses a combination of quantitative, qualitative, and formal theoretical methods to examine the mechanisms that link social inequality to the health of populations, individuals, and the human body more generally. Corey currently manages UC Berkeley's interdisciplinary* **Center for Urban Ethnography (CUE)** *where he collaborates on projects with scholars from various fields, works on methodological issues, and directs an undergraduate training program in qualitative methods. He is also the lead instructor in Computer Assisted Qualitative Data Analysis (CAQDAS) for the* **UCB Social Research workshops***, where he provides methodological training and consulting to faculty, graduate students, and research teams from across the country.*

Introduction

When I teach workshops on data analysis in ATLAS.ti, I get to interact with people from a wide spectrum of disciplines and philosophical orientations. I consider this ability to accommodate a host of methodological and epistemic perspectives to be an important strength of current generation Computer Assisted Qualitative Data Analysis Software (CAQDAS). The software accommodates phenomenologists, interpretivists, grounded theorists, positivists, and a whole host of other epistemic orientations. ATLAS.ti does not require the analyst or team to impose an exclusively top down-deductive logic, or an emergent inductive logic, but allows the deployment of both logics, either alone or in combination. Because of this flexibility, CAQDAS workshops are one of the few venues where it is not odd to have political scientists, sociologists, public health scholars, rhetoricians, anthropologists, and market researchers in the same room. Often, the key common denominator is that they recognize the software as useful for their research or inquiry (even if they agree on little else). Consequently, when I teach ATLAS.ti I do not proselytize or try to advocate for a particular methodology or epistemic perspective, but rather show how the software can be used as a flexible tool to accomplish individuals' particular goals. I always believe that the "best practices" for using software, like any tool for social research, are those practices that work best given the user's philosophy and aims.

Nonetheless, as a sociologist that does work on inequality, health, and policy at a large research university, I operate within a particular scholarly and organizational context. That context requires that I contend with classical positivist-behavioral concerns about bias, error, representation, inferential logic, and reliability. Many of the individuals and teams I work with use ATLAS.ti in similar organizational and disciplinary contexts. Some agree with the underlying logic of positivist-behavior research, others find it repugnant. There are myriad critiques and responses (some of which I list at the end of this article). Still, regardless of where they fall in these debates, researchers acknowledge that there is a comparative dearth of literature relating CAQDAS to these concerns. This article is not a normative assessment of various epistemologies or the conditions under which one might adhere or deviate from them. My intent is not to provide prescriptive statements about what constitutes "proper" research. Rather, I address (in a limited way) how CAQDAS relates to concerns with the possibility of error in inferred associations under the positivist-behavioral tradition, and show how some existing tools in ATLAS.ti can be used by those operating in this tradition.

CAQDAS and the Problem of Inference

Two of the most common questions I get asked during workshops and consulting are: 1. "How do I know if an association between codes is real?" and 2. "How do I deal with relationships that are real, but do not show up in my examination of code associations?" Both questions are concerned with the representation of a social or linguistic process in a data set, and how it maps on (or fails to map on) to a reality outside. It is a component of a set of classical problems in analytical inference- how does one get correspondence between recorded data and the phenomena they purport to record? How can theory or model be made line up with the phenomena they explain? And on a related note, how does one minimize the possibility of error, or miss-correspondence? In essence, the "how do we know" questions are complex philosophical problems underwritten by particular logics of inference. In the positive-behavior tradition, the overarching concern is how can we deal with the lurking specter of type I errors (false positives) and type II errors (false negatives) in research generally – and qualitative research more specifically.

Dealing with problems and theories of inference and correspondence is a huge field, with volumes of books and articles addressing possible responses. My contribution in this article is much more meager. I do three things: (1) I address how one might examine the relationship between two or more "codes," (2) I examine how this process relates to the concern of type I and type II errors, and (3) I discuss some of the new tools in ATLAS.ti that can be deployed to help minimize inferential errors.

Defining Codes and Coding

Before I examine how to examine the relationship between codes, it is necessary to articulate what social scientists in the positive-behavioral tradition typically understand as codes and coding in qualitative research. To put it generally, coding is the process by which an investigator or team marks specific pieces of data (text, images, audio clips, video, geo-spatial locations) as instances of particular categories, events, concepts, or variables. In ATLAS.ti, codes break documents into a smaller unit of data known as a "quotation." The code is an indicator that a piece of data (the quotation) contains characteristics that are part of some larger grouping. Codes can range from seemingly straightforward demographic constructs (e.g. age: 20-29, state:california), to abstract theoretical notions such as "cultural capital," "anomie," or "collective effervescence," to the mezzo level constructs in between (e.g. social network types, attitudes, various cultural tropes etc.) ATLAS.ti offers numerous ways to arrange these codes into larger hierarchical and non-hierarchical groupings, and to search for quotations where codes or combinations of codes co-occur (or overlap). Codes are flexible and non-exclusive. Many codes can be applied to one bit of data, or codes can be created that apply to no data at all. By applying a code to a piece of data, the analyst is essentially saying "this bit of data is an instance of [something]." How this something is measured, understood, generated, or "operationalized", is ultimately chosen by the researcher. In essence, all coding is an analytical imputation. It is saying this piece of data (1) should be considered/understood as a piece of data and (2) is somehow associated with the code and its larger categories. Even when code names are generated using respondent speech behavior (as in invivo coding), there is an analytical moment whereby the researcher designates a portion of the raw data as something to be analyzed, and denotes the speech behavior itself as the code with which it is tagged. Before CAQDAS software, codes were often simply keywords written in the margin of interview transcripts or fieldnotes, newspapers, pamphlets, maps, etc. At the most basic level coding assists the analyst in understanding the massive amount of qualitative data by referencing and cross-referencing key observations.

There are two common ways of generating codes in qualitative social-science research. The first is the inductive or emergent generation of codes. Most "qualitative methods" such as participant observation, content analysis, and in-depth interviewing typically have inductive components. That is to say, emerging previously unforeseen patterns/occurrences in the data form a part of the analysis. Inductive codes, generated after the analysis or fieldwork has begun, are used to tag and/or group these events/quotes into meaningful categories. These codes are often then grouped into even larger meta-categories, which facilitate the production of a model, theory, or explanation. Some methodologies (e.g. grounded theory) use this form of coding exclusively, others verbally discard it as subjective. In practice, the inductive generation of codes is a common aspect of most qualitative research even in the most rigidly positive-behavioral disciplines, since to ignore unforeseen patterns would be to miss out on one of the key strengths of qualitative research.

The second mode of generating codes is deductive. This sort of coding is common for positivist behavioral research, clinical methodologies, the extended case method, and numerous other approaches found in social science and policy disciplines. Here, researchers begin a project with an understanding of specific existing theories that purport to explain why things happen the way they do in the world. They want to account for these in their analysis of qualitative data. After doing a project/grant proposal, literature review, etc., these researchers will typically go through and generate codes based on existing, possible, and counter-factual explanations for the social phenomenon they are studying. I refer to these codes, which are typically generated before fieldwork or data analysis as deductive codes. Deductive codes operationalize existing theories, explanations, or categories from prior empirical and theoretical inquiry. As with inductive codes these can include everything from demographic categories such as age, to abstract theoretical constructs. The codes may not correspond to, or be applied to the data, but they exist initially to orient a researcher to an analytical frame. If they correspond to the data (even in part), that is a key finding. If they do not correspond at all (e.g. they are useless in explaining a social phenomenon), that is a finding as well. Generating at least some codes deductively also saves a great deal of time in the process of recoding.

The most common practice in sociology and related disciplines is currently to use a combination of inductive and deductive codes. Deductive codes are generated based on prior understandings of a topic, existing theories, and hypothesized explanations that purport to explain a specific research puzzle. Inductive codes are generated to correspond to the unforeseen patterns and occurrences that present themselves during the course of producing and analyzing qualitative data. Since, unlike survey methodologies, the range of "responses" or observations in qualitative research are not limited to a drop down menu pre-generated at the outset of a project, inductive codes are necessary to account for observed occurrences or patterns that were not in the initial analytic frame. Deductive codes are generated using a top down approach- e.g. starting with explanations and seeing if they map on to data. Inductive codes are based on a bottom up approach- beginning with the data and building up. The extent to which one uses inductive versus deductive codes (or even uses codes) is a function of their methodological tradition and epistemic choices.

Codes, Errors, and Inference

There are two types of relationships between codes (key concepts, themes, or variables) in a qualitative data set. First, there is a mathematical/formal relationship between the codes in ATLAS.ti. This formal relationship determines what comes up in the window when one does a query or produces advanced output like co-occurency tables. These codes can be related by boolean links, semantic links, or proximity links. The ATLAS.ti manual's discussion of the query tool explains these links in detail. In short, these links allow the researcher to see if codes come up together, come up alone, or are following/preceding/overlapping one another. There are more complex associations that are possible, but that is the basic idea.

The second type of relationship is a substantive empirical relationship. Some of the formal associations that show up in ATLAS output may indicate a spurious relationship. The classic example in elementary statistics is the positive correlation between ice cream consumption and drowning. The common argument here is that the association is spurious, explained by an unmeasured explanatory variable – time of year. Both are explained by the coming of the summer months, which increases both ice cream consumption and swimming. If we agree to this sort of logic, the association between ice-cream and swimming deaths is a statistical artifact. It is what can be a called a type I error or false positive. There is a formal/mathematical association between two concepts or codes, but the association is ultimately spurious (or non-substantive) and explained by something else- another measured or unmeasured variable, concept, or code (in this case time of year).

It is also possible that two concepts or the codes that represent them are related in a crucial way, but that codes do not come up as formal associations in queries. This is referred to as a type II error of false negative. In this instance, there is a real relationship between two things in the world (e.g. time of year, swimming, and likelihood of drowning), but this relationship is not easily reflected in the data. Using the example above one might search for time of year-summer and drowning, but the resulting query does not yield useful results. Or perhaps, there is no existing code or concept to refer to what is assumed to be the explanatory factor (time of year). Here, the lack of formal association in the software does not map onto the substantive association between time of year and drowning that exists in the larger universe of investigation. In addition to the type one error (the false positive), the apparent lack of a relationship-when one exists in the universe outside- is a type two error (a false negative).

I am often asked questions that knowingly or unknowingly touch on this issue of inference. People ask "how do I make sure my data and coding capture what is going on?" "how do I make sure a relationship isn't just made up?" and "what if there is something I know is going on in the world, but isn't showing up in my data set?" Often, the underlying assumption is that there is a simple answer and that their difficulty is the result of a lack of technical proficiency with ATLAS.ti. The truth is that while technical proficiency is necessary to address these problems, particularly for large qualitative data sets, these are complex methodological questions. They speak to the palpable tension of how to balance type I and type II errors in the construction and analysis of data. There is no single command (i.e. right-click- minimize type I errors) or technique that will dissolve this problem. There is no one-size-fits all solution. Even within the methodological logic outlined above (which is only one of many responses to epistemological issues such as how to manage “truth” and “correspondence”), the researcher and analyst must make hard decisions about how to balance and manage type I and type II errors in specific projects. While there is no one size fits all solution, in grappling with these issues in both my own research and that of others, I have developed some thoughts and opinions that may be of use to others.

Dealing With and Minimizing Error

My general opinion is that type I errors (false positives) are a lot easier to rule out in qualitative research than in quantitative research. A qualitative researcher can always go back and investigate a potential relationship by rereading interviews etc, looking at a text, listening to audio, then reassessing if the relationship is spurious based on a re-analysis of the initial content. In primarily quantitative research, false positives are harder to deal with, since the assessment is typically made by examining the numeric strength of association between measures that are largely set at the outset of a project. Some recoding is possible (e.g. collapsing an interval variable into an ordinal variable, or building up binary variables into an index), but the initial categories involved in data collection and analysis are relatively set. Without additional data (e.g. a record of the observed event), it is hard to rule out the possibility of miscoding or mis-imputation. Qualitative methods furnish additional information for ruling out false positives due to errors in the coding scheme, a deficient initial analytical scheme, or chance overlap. They do so, by providing the analyst to look back at the data on which a coding association is based (e.g. seeing if eating ice-cream has an effect on swimming prowess), and examine whether that apparent relationship is an artifact of methodological decisions or mis-imputation. Ultimately, this assessment comes back to the researcher and their scholarly community, but in the case of qualitative research, the raw data provides an additional tool for adjudication.

Type II errors (false negatives) are harder to deal with, because they are invisible. Broadly, this is an issue of missing something either in the observation of a social phenomenon or its analysis. Returning to the issue of code associations, type II errors refer to real relationships that don‘t come up as an association in queries or output. In statistical methods, type II errors are often cast as less dangerous, since no association is seen as the “more conservative” result. But in reality, they are still errors. The analyst misses part of what is going on in the data set and the world it is meant to represent. Unfortunately, type II errors are harder to rule out, because the analyst may not even know they are there. The ability to generate codes inductively and reanalyze data in qualitative research assists in this capacity (again by allowing the analyst to return to pre-coded data and reformulate categories), but it does not remove the problem.

My strategies for minimizing error in qualitative data analysis rely on the fact that it is easier to rule out false positives than false negatives in qualitative research. I try to minimize type II errors by generating codes deductively at first, creating codes inductively to correspond to new findings, coding densely with all relevant codes, then going back (using the query tool) to rule out false positives. I look at formal associations as something potentially true. I then decide if this is a false positive, not by computing the probability of getting such a result due to chance, but re-examining the event or phrase (in ATLAS.ti typically the quotation) where the codes overlap and examining the quality of the relationship.

Ruling out false negatives is harder, and requires more cognitive flexibility. It often involves re-reading or searching through text in a less directed way. If there is a pattern in the data that emerges, but is not found in an existing code or codes, I create a new code or codes so I can reference it easily. If I can convince myself (and hopefully the reader) that a substantive relationship (1) exists and (2) refers to something real in the social world, I include it in my analysis and say that the failure of my initial coding scheme to capture this is the result of a type II error. Since the substantive claim is not based on a probability of error, but on an observed and demonstrable pattern or occurrence, this is consistent with the logic of inference espoused above (this is bracketing the issue of statistical induction—or inference from sample to population). It is a strength of qualitative research to account for and incorporate observations the analyst didn‘t expect. The ultimate concern isn‘t simply how good the initial coding scheme is, but whether or not the emerging analysis, which relates to both existing theories and the real world is correct in the end.

Tools for Reducing Type II Errors in ATLAS.ti

Given this general strategy for dealing with error, there are a number of basic and advanced tools in ATLAS.ti that

I use to this end. Each of these are covered well by existing articles and documentation in ATLAS.ti. The idea is to think about these not simply as strategies for exploring data, but for decreasing type II errors. Each of these can be used to help patterns you may have missed.

- Rereading data.

- Control-f allows you to search for strings of text (in a way that may not line up with your intial coding).

- The object crawler does this as well, and provides more options.

- In the query tool you can use the "follows" and "precedes" operators instead of limiting yourself to Boolean (and, or, not, Xor) searches.. This can help you in identifying chronological patterns.

- Using quantitative outputs—e.g. the code primary-document table could broaden the scope in looking for a relationship between codes. You could use this to see for instance, if and how many times two codes are in same primary document.

- Co-occurency table explorer. This is a new and extremely powerful tool for seeing patterns in your data. See the new newsletter article by Ricardo B. Contreras here: **http://downloads.atlasti.com/library/contreras_nl201108.pdf**

- You can use network views, to examine the association of codes from different vantage points.

Conclusion

The questions "how do we know if an association is real," and "how do we deal with real associations that do not show up in a query," are not technical questions. They are fundamental issues tied up with epistemological concern like "how do we know what we know." As such, there is no simple answer, and the most useful or palatable solution varies between and within disciplines. While criticisms of the positivist-behavioral tradition abound-particularly in qualitative research, this is context or tradition in which many of us operate. Without offering prescriptive statements about the relative utility of different epistemologies, I have discussed a problem fundamental to research in the positivist-behavioral tradition: dealing with and minimizing errors in association. I showed how this relates to CAQDAS, and the specific tools that ATLAS.ti has for this purpose. It is my hope that even if one wants to contest this logic, having a better understanding of it, and the way ATLAS.ti's tools might be implemented, can strengthen their own inquires.