Article

Sociological Methods & Research
42(3) 294-320

DOI: 10.1177/0049124113500475
smr.sagepub.com

Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement

John L. Campbell[1,2], Charles Quincy[3], Jordan Osserman[4], and Ove K. Pedersen[2]

Abstract

Many social science studies are based on coded in-depth semistructured interview transcripts. But researchers rarely report or discuss coding reliability in this work. Nor is there much literature on the subject for this type of data. This article presents a procedure for developing coding schemes for such data. It involves standardizing the units of text on which coders work and then improving the coding scheme's discriminant capability (i.e., reducing coding errors) to an acceptable point as indicated by measures of either intercoder reliability or intercoder agreement. This approach is especially useful for situations where a single knowledgeable coder will code all the transcripts once the coding scheme has been established. This approach can also be used with other types of qualitative data and in other circumstances.

[1] Department of Sociology, Dartmouth College, Hanover, NH, USA
[2] Department of Business and Politics, Copenhagen Business School, Frederiksberg, Denmark
[3] IBM Corporation, New York, NY, USA
[4] Birkbeck College, University of London, London, United Kingdom

Corresponding Author:
John L. Campbell, Department of Sociology, Dartmouth College, 123 Silsby Hall, Hanover, NH 03755, USA.
Email: john.l.campbell@dartmouth.edu

Keywords

qualitative methods, coding, intercoder reliability, intercoder agreement, interviews

In-depth semistructured interview data constitute the empirical backbone of much qualitative research in the social sciences. This includes, for example, seminal studies in such diverse areas as urban inequality (Wilson 1996), gender studies (Edin and Lein 1997; Gerson 1985), organizational sociology (Strang 2010), and economic sociology (Halliday and Carruthers 2009; Useem 1984). What concerns us here is how one establishes the reliability of coding for this type of data and especially when only one person is doing the coding. This is a problem insofar as studies using this type of data are done by researchers whose budgets do not permit hiring a team of coders and who are the only principal investigator (PI) as is often the case, for example, with young researchers just starting their careers. In fact, several studies that fit this profile have been published recently by young researchers in top sociology and other journals (e.g., Calarco 2011; DeSoucey 2010; McCabe 2011). This article offers a methodology for developing reliable code schemes for this type of data under these conditions.

Reliability is just as important for qualitative research as it is for quantitative research (Becker 1970; Deutscher 1970; Zelditch 1970). However, the issue of reliability is discussed more often in research studies based on quantitative rather than qualitative data. Hence, some researchers advocate confronting the issue more explicitly in qualitative work. In particular, they call for more attention to determining whether qualitative measurement instruments provide consistent results across different coders, raters, or observers (Popping 2010:1068; Tashakkori and Teddlie 1998:75-85). This should be possible insofar as the same methodological and epistemological principles that are applied to quantitative research can also be applied to qualitative research (Greene, Caracelli, and Graham 1989).

As is well known, there are three types of reliability (Krippendorff 2004:chap. 11). One is *stability* where the concern is whether a coder's use of codes changes over time. Second is *accuracy* where a gold standard coding scheme is already established with high reliability and other coding schemes are developed and compared to it. Third is *reproducibility* across coders—often called *intercoder reliability*—where the concern is whether different coders would code the same data the same way. Our interest here is with developing coding schemes that are reproducible for in-depth semistructured interviews.

Much has been written about intercoder reliability but surprisingly little has been written about it for coding text based on in-depth semistructured interviews (but see, Bernard 2000:chap. 11; Hruschka et al. 2004; Kurasaki 2000).[1] Prominent researchers using interview data do not generally report in their publications whether they assessed the reliability of their coding. And if they do report the level of intercoder reliability, they typically do not explain how they determined it (e.g., Edin and Kefalas 2005; Edin and Lein 1997; England and Edin 2007; Gerson 1985; Wilson 1996).[2] For example, Lombard, Snyder-Duch, and Bracken (2002) reviewed 137 research articles based on content analysis of various sorts including interviews. Only 69 percent of these articles contained any information on intercoder reliability. And in most cases, the information provided was extremely sketchy and ambiguous. Others have made similar observations and lamented this oversight as well (Fahy 2001; Hruschka et al. 2004; Riffe and Freitag 1997). This omission may stem from the fact that some qualitative researchers using text data are skeptical of reliability tests because their data do not take the form of a clearly standardized set of measurements (Armstrong et al. 1997; Mason 2002:187). This is due to the fact that words may have multiple meanings, may be open to interpretation, and may only be understood in the context of other words, which in a way makes them harder to work with than numbers (Miles and Huberman 1984:54). Indeed, there are few agreed-on canons in this regard for qualitative data analysis (Armstrong et al. 1997; Miles and Huberman 1984:chap. 1).

Moreover, the literature on establishing reliable coding schemes for qualitative content analysis does not focus on interviews but other types of data like field notes (Miles and Huberman 1984), documents (Krippendorff 2004; Weber 1990), conference discussion transcripts (Fahy 2001; Garrison et al. 2006), ethnographies (Hodson 1999), and observations of behavior recorded on videotape (Rosenthal 1987). Even extended discussions of interviewing barely address the subject (e.g., Garrison et al. 2006:2; Mason 2002:chap. 4; Rubin and Rubin 2005). There is a small literature on how reliable survey data are in comparison to coded interview data, but it still begs the question of how to establish coding reliability for these transcripts in the first place (e.g., Engel 2007).

To be sure, there are some basic lessons to be gleaned from this literature that can be applied to coding in-depth semistructured interview data, such as defining codes clearly and in mutually exclusive ways in order to enhance intercoder reliability. But in-depth semistructured interviews present some unique challenges in this regard. For example, unlike the coding of ethnographies, which take the entire ethnography as the unit of analysis (e.g., Hodson

1999), in-depth interviews often involve many units of analysis, which are not always easily identified. As will be discussed later, this *unitization problem*—that is, identifying appropriate blocks of text for a particular code or codes—can wreak havoc when researchers try to establish intercoder reliability for in-depth semistructured interviews. And unlike tightly structured interview questionnaires, which tend to elicit short responses requiring only a single code for each one, semistructured interviews tend to elicit more open-ended, rambling responses that often require several codes simultaneously.

In short, there is not much guidance in the literature for researchers concerned with establishing reliable coding of in-depth semistructured interview transcripts. And there is virtually none for establishing reliability in the situation where coding is left up to a single coder, particularly one who needs to be knowledgeable enough about the subject matter in question to identify subtle meanings in the text. The need for knowledgeable coders is especially important when working with in-depth semistructured interviews. Coding this type of data often involves interpreting what respondents mean in their answers to questions. Doing so correctly requires that coders have sufficient background knowledge in the subject matter of the interviews.

This article offers a methodology for establishing intercoder reliability and intercoder agreement for coded in-depth semistructured interview transcripts. The goal is to ensure that a single knowledgeable coder may be reasonably confident that his or her coding would be reproducible by other equally knowledgeable coders if they were available. The difference between intercoder reliability and agreement is important. *Intercoder reliability* requires that two or more equally capable coders operating in isolation from each other select the same code for the same unit of text (Krippendorff 2004:217; Popping 2010:1069). *Intercoder agreement* requires that two or more coders are able to reconcile through discussion whatever coding discrepancies they may have for the same unit of text—discrepancies that may arise, for instance, if some coders are more knowledgeable than others about the interview subject matter (Garrison et al. 2006; Morrissey 1974:214-15). Generally speaking, the issue of intercoder reliability is discussed frequently in the methods literature but the issue of intercoder agreement is not.

The contributions of this article are fourfold. First, it brings together and shows how methods for establishing intercoder reliability and agreement can be utilized in a very practical way to improve the quality of coding in-depth semistructured interview transcripts when there is only one knowledgeable coder available. Second, it shows how the problem of unitizing transcript text (i.e., identifying appropriate blocks of text for coding) complicates the

process of establishing intercoder reliability and agreement. Third, it provides a solution to the unitization problem for this type of data. Finally, it discusses how these methods might also be applicable in other research situations.

As such, this article helps fill a significant void in the qualitative data analysis literature, which has neglected an important and popular type of qualitative data. We offer this as a starting point that we hope will encourage researchers to be more explicit about how they develop reliable coding schemes for this type of data. In doing so, we draw on the extant literature on qualitative coding wherever possible. We argue that coding in-depth semistructured interviews should typically involve a three-stage process. The first stage involves developing a coding scheme with as high a level of intercoder reliability as possible based on a sample of transcripts. The second stage involves adjudicating the remaining coding disagreements through a negotiation among coders in an effort to establish a high level of intercoder agreement. The third stage involves deploying that coding scheme on the full set of transcripts once acceptable levels of intercoder reliability and/or agreement have been achieved. Two coders are required for the first two stages. The third stage requires only one knowledgeable coder. Our focus here is on the first two stages.

Let us be clear about two things. First, just because researchers often fail to report how they coded in-depth semistructured interview transcripts does not necessarily mean that they have failed to develop a coding scheme with care or that they have failed to check its reliability. Second, the approach we present for developing reliable code schemes for this type of data is not the only one on offer. We discovered all this during a revision of this article during which we contacted other researchers who have worked with this type of data. We report some of their approaches later. But the important point for now is that there is very limited guidance on these matters in the literature. Indeed, we hope that this article will encourage others to consider these issues at greater length and contribute their own insights to the literature on the subject.

Relying on our own research project for illustrative purposes, we proceed as follows. First, we provide a brief overview of how we developed our coding scheme. Second, getting to the heart of the matter, we discuss how we assessed and improved the coding scheme's reproducibility as we developed it. This involved solving two problems. One was devising a way to *unitize* the text so that all coders were coding the same sections of text—something that was necessary to facilitate accurate comparisons of each one's coding. The other was improving the coding scheme's *discriminant capability*—that is,

reducing coding errors. This was especially tricky because the PI, who would be responsible for much of the coding later, was far more knowledgeable about the interview subject matter than his assistant coder. Hence, in addition to assessing intercoder reliability it was necessary to assess intercoder agreement. Third, we review our method of calculating the coding scheme's intercoder reliability and intercoder agreement. Fourth, we briefly discuss some other approaches that we learned about from discussing these issues with other researchers. Finally, we conclude with a discussion of the limitations and possible further applications of our approach.

The text data to which we refer throughout this article consisted of 101 in-depth semistructured interviews conducted with people in policy research organizations, such as think tanks and government research units, for a research project on economic policy analysis and advising (Campbell and Pedersen 2014). Interviews averaged 90 minutes in length and were recorded digitally. The recordings were then transcribed verbatim as Word documents. Transcripts were typically between 20 and 30 single-spaced pages long. These were loaded into Atlas.ti (version 6.2) qualitative data analysis software, which we used to develop our coding scheme and then to code the transcripts. One need not use Atlas.ti. There are several good qualitative data analysis software packages available that should suffice (e.g., Lewins and Silver 2007). The purpose of this coding scheme was not to quantify the data for statistical analysis but to ensure that the PI could later retrieve all portions of the transcripts that pertained to particular coded themes without omitting any relevant portions that should have been coded. Coding focused on different themes discussed by respondents in the interviews, such as how their organizations did policy analysis, how the organization changed over time, why it changed, and how respondents thought their organization influenced public policy making.

Developing the Coding Scheme: A Brief Overview

We developed our coding scheme in ways that were as consistent as possible with the extant literature. The most basic steps are described in this section of the article. They are well known and typical of much code development for quantitative and qualitative research. Hence, we will be brief. However, there is no simple or right way to do this sort of analysis, so it must be tailored to the project (Weber 1990:13, 69). In our case, innovations were required that we present in more detail later.

To begin with, the PI hired an undergraduate research assistant to help develop the coding scheme. He trained him in Atlas.ti, had him read a detailed

description of the project to become familiar with it, and explained the purpose of coding. The PI developed the first version of the coding scheme beginning with an initial set of several dozen codes with definitions grouped into categories. He then discussed the coding scheme with the research assistant explaining how it was developed, what the codes were, and what each one meant. Based on this discussion, they clarified the codes and definitions.

The PI then selected at random an interview transcript. We each coded a copy of it. Then, we compared results and discussed coding problems where there were discrepancies and confusion. We discovered that the coding scheme was too cumbersome, complicated, and sometimes confusing and that intercoder reliability was low, so we revised the codes again. Adjustments like this are not unusual (Hruschka et al. 2004; Miles and Huberman 1984:56-60). A discussion of how intercoder reliability (and later intercoder agreement) was assessed and calculated comes later in the article. To improve the situation, initially we followed those who advise that the appropriate response to low intercoder reliability is to drop or merge unreliable codes in order to reduce the number of codes that coders need to remember, clarify coding definitions, modify codes and/or coding instructions, repeat the reliability test, and then repeat the exercise until an acceptable level of reliability is achieved (Hodson 1999:24-26; Hruschka et al. 2004). We did this several times devising and writing in a codebook ever more precise and concrete code definitions with examples where necessary. (See online appendix e.g., from the codebook which can be found at http://smr.sagepub.com/supplemental/.)

It is acceptable to assess intercoder reliability on a sample of the texts to be analyzed, especially when costs prohibit multiple codings of every text (Krippendorff 2004:chap. 11). However, there is little agreement as to how large a sample of texts is appropriate. Some recommend using 10 percent of the set of documents (Hodson 1999:29), while others argue that as few as 5 to 10 pages of one set of transcribed field notes is sufficient (Miles and Huberman 1984:63). We began with one full-length interview transcript. In every subsequent coding trial, we used a different full-length interview transcript. In the end, we used 11 transcripts before we were satisfied with the results—roughly 10 percent of our transcripts. The implication is that researchers should continue to sample transcripts and refine the code scheme until they are satisfied with the level of intercoder reliability (or agreement).[3] We discuss below what constitutes a numerically satisfactory threshold of agreement among coders.

There is a fine line between pressing for simpler coding schemes—notably by limiting the number of codes—and remaining sensitive to nuance in

the data (Hruschka et al. 2004:325). So in trying to improve the coding scheme, we tried not to sacrifice meaning by simplifying it to the point that intercoder reliability (or agreement) was high, but the codes did not capture much that was important for the project. Put differently, reducing the number of codes in order to achieve a satisfactory level of reliability should not be taken to such an extreme that the coding scheme cannot be useful for identifying or differentiating among important nuanced meanings in the data.

We eventually developed a coding scheme based on *code families*—that is, several codes reflecting different aspects of a general theme. An example is the "research" code family presented in online appendix (which can be found at http://smr.sagepub.com/supplemental/) that includes a *primary code* for all text referring to the research processes of the organizations we interviewed and then several *secondary codes* referring to more detailed aspects of their research processes, such as the methodologies they used, whether they outsourced research to external experts, and so on. Our final coding scheme was more complex than most others described in the literature. It included 60 codes (9 primary and 51 secondary) excluding the names of organizations.[4] It had a moderate level of intercoder reliability and a high level of intercoder agreement.

Assessing Intercoder Reliability and Agreement

The evaluation of intercoder reliability and agreement should be part of the development of coding schemes for qualitative data in order to satisfy people that the data are sound (Hruschka et al. 2004; Krippendorff 2004:chap. 1; Miles and Huberman 1984; Weber 1990). But assessing reliability and agreement for in-depth semistructured interview data involves several challenges. We confronted two that required us to be innovative.

The first involved the *discriminant capability* of the coding scheme. This involves determining how well coders could readily and unambiguously categorize text content (Fahy 2001; Kurasaki 2000). Researchers should worry about this especially insofar as their code schemes are complex as ours was. Complex coding schemes are less reliable than simple ones (Garrison et al. 2006). The more codes a coder must keep track of at one time, the more cognitively difficult coding becomes and the greater the chance for coding errors (Fahy 2001; Hodson 1999:24-26; Hruschka et al. 2004:319). As discussed below, this is why we eventually began coding each transcript twice—once with half the codes and then again with the rest. The complexity of coding was exacerbated for us by the fact that there were often several codes that could be applied to the same section of text, which is common

with this type of data. In our case, a paragraph might contain simultaneous discussions of several different issues, such as change in an organization, the causes of that change, and the effect of that change on an organization's influence on the policy-making process.[5]

The second problem involved *unitization* (Krippendorff 1995). This problem arises when portions of text to be coded—the units of analysis—are not naturally given, such as a sentence or paragraph, but require the subjective interpretation of the coder. This is a problem insofar as different coders may unitize the same text differently. Why? Because they may disagree on which segments of text contain a particular meaning (Kurasaki 2000). For example, two coders may each identify a string of text for the same code. But while each of their strings may overlap they may still vary in length because one coder includes text providing background information that helps establish the context for the code in question but the other coder does not (Fahy 2001; Krippendorff 1995; Kurasaki 2000). Particularly in situations like ours where we coded free responses to open-ended questions, identifying the appropriate unit of analysis can be difficult. This makes it hard to assess intercoder reliability and agreement. If coders do not unitize a text exactly the same way, it may become difficult to determine whether their coding is the same. In other words, one must first solve the unitization problem in order to assess the discriminant capability of the coding scheme.

Solving the Unitization Problem With Units of Meaning

There is considerable debate about whether clearly demarcated parts of the text, such as a sentence or paragraph, rather than "units of meaning" as defined by the coder are the appropriate units of analysis (Garrison et al. 2006; Morrissey 1974). On one hand, the concern with using predefined blocks of text is that it may not accurately reflect the meaning as intended by the respondent. On the other hand, the concern with using a unit of meaning is that it involves coder subjectivity where the demarcation of any meaning unit depends largely on how the coder interprets it (Fahy 2001; Kurasaki 2000). In exploratory research and research using complex interview data, such as ours, the meaning unit may be the appropriate unit of analysis because it is less likely to decontextualize what the respondent is saying (Garrison et al. 2006). When meaning units are used, the unitization problem is compounded if respondents do not speak in clearly delineated text units. Indeed, data like our in-depth semistructured interview transcripts are conversational and choppy. Interviewers may ask about a particular theme at certain moments but frequently receive answers that include

tangents, digressions, backtracks, and overlaps with other themes (Kurasaki 2000:180).

The point is that there are no clear guidelines in the literature as to what the appropriate unit of analysis should be for coding interview transcripts (Hruschka et al. 2004:311; Kurasaki 2000). Notably, in one of the few extended discussions of unitization, Krippendorff (2004:chap. 5) reviews the advantages and disadvantages of unitizing text of various sorts in different ways but offers virtually no guidance on which approach is most desirable for in-depth semistructured interviews. Nor does he say who should do the unitizing. In his words, this choice, "poses many epistemological questions that I cannot attempt to address here." In fact, the only advice he offers is to suggest that, "This act [of unitizing text] depends on the analyst's ability to see meaningful conceptual breaks in the continuity of his or her reading experiences, on the purposes of the chosen research project, and on the demands made by the analytical techniques available to date" (p. 98). We will return to this advice later. But remember that for Krippendorff the level of coder knowledge is important here.

In our case, the transcripts ran on for many pages, covered many themes and, as a result, were too long to take as a single unit of analysis. Sentences or single paragraphs were often too short to capture the full meaning of what a respondent was saying about a particular theme. Besides, responses to our questions varied from a few sentences to several paragraphs. It is not uncommon for analysts to simply take the transcript page as the unit of analysis for calculating intercoder reliability (Bernard 2000:46-62). But taking a transcript page as the unit of analysis was unsatisfactory in part because Atlas.ti presents a transcript as one continuous flow of paragraphs without page breaks. Moreover, coders used their own computers with computer screens of different sizes. So the length of a transcript page varied across coders according to the size of their screens.

For these reasons and because we were engaged in exploratory research with complex interview data, we pursued a unitization strategy that focused on meaning units rather than naturally given units of some sort. Coders marked as a codable unit any portion of text regardless of length to which they believed a code applied. Some coded units were a sentence or two, some were a full paragraph, and some were more than a paragraph including sometimes the end and beginning portions of two adjacent paragraphs if not more. However, because coders unitized the same text in different ways it was difficult to compare their codings to determine intercoder reliability or agreement—that is, the discriminant capability of the coding scheme. We ran into this problem immediately.

To solve the unitization problem, we devised the following procedure. The PI identified the meaningful units of analysis. In order to do this, he coded an interview transcript with Atlas.ti. This involved marking a segment of text in the margin with a bracket and then placing the appropriate code/codes alongside the bracket. Once the text had been fully coded he saved it on his computer. Then on a copy he removed all the codes but not the brackets. He then gave the bracketed—but no longer coded—version to his research assistant who then coded the already bracketed sections. In this way, both coders coded exactly the same units of text. Then, they compared their coded transcripts to see whether there were discrepancies in their coding.

The reason for having the PI do the unitizing is straightforward. Unitizing and coding text in this situation requires subjective interpretation, contextualization, and especially a thorough understanding of the theoretically motivated questions guiding the study. Recall Krippendorff's advice now. The ability to see meaningful conceptual breaks depends very much on the qualifications of the coder and his ability to discern not only obvious meanings, such as specific words, phrases, or organizational names, but also more subtle meanings of a respondent's statements.[6] The PI's qualifications in this regard were greater than his assistant's because he was far more knowledgeable about the interview subject matter than the assistant was. In any case, the important point is that unitizing the text in this way eliminates a potential source of confusion when comparing the coding of two or more coders especially when one is more knowledgeable than the rest. Thus, this approach to unitizing text facilitates a more accurate assessment of intercoder reliability.

One might wonder whether solving the unitization problem like this might inflate or otherwise bias the level of intercoder reliability or agreement. Might not unitizing the text in this way alert coders to the fact that there is something to be coded in a portion of the text that they might not have coded otherwise? And might not unitizing the text in this way lead coders to focus more on some codes than others? Perhaps. We address this issue later. But what is important now is that we were not particularly concerned about this in our case because the problems we discovered in the first place that led us to unitize transcripts the way we did were often about how much text to bracket for a particular code rather than whether a particular code was appropriate. In particular, one person tended to code a single paragraph with a particular code, whereas the other person tended to code that paragraph as well the two or three that preceded it with the same code in order to include more contextual or background information.

In any case, intercoder reliability improved with our unitizing procedure but was still not satisfactory. Through an iterative process of unitizing,

coding, discussing coding discrepancies, and refining codes and code definitions intercoder reliability gradually improved and finally reached a plateau of 54 percent reliability on average for the primary and secondary codes combined and 65 percent reliability for the primary codes alone. These results were somewhat disappointing but not necessarily surprising, given this type of research. First, exploratory studies, such as ours, where no already existing and reliable coding scheme is available to emulate should anticipate lower levels of reliability than studies where proven coding schemes already exist (Krippendorff 2004:chap. 11). Second, coding paragraphs and longer passages like we did tends to reduce reliability relative to coding sentences or a few words because there is more room for coders to interpret what is being said, and because more than one theme and therefore more than one code may be applicable at once (Weber 1990:39). Only very rarely did a unit of text in our work receive a single code. Third, as noted earlier, our coding scheme was quite complex. The more codes one needs to keep track of the greater the chance there is for coding errors (Fahy 2001; Hodson 1999:24-26). This problem is compounded for in-depth semistructured interviews like ours that are more likely to elicit wide-ranging and even rambling answers from respondents than more closely structured interview protocols, which tend to have higher levels of intercoder reliability (Hruschka et al. 2004; Rosenthal 1987:81-83).

We suspected, however, that differences in the level of knowledge that each coder brought to the text during coding may have been responsible as well for the fact that intercoder reliability was not as good as we wanted. After all, the PI was much more knowledgeable about the project than the assistant. To address this issue, we turned to a procedure for assessing intercoder agreement.

Solving the Discriminant Capability Problem With Negotiated Agreement

Some scholars have adopted a ''negotiated agreement'' approach for assessing intercoder reliability where two or more researchers code a transcript, compare codings, and then discuss their disagreements in an effort to reconcile them and arrive at a final version in which as many discrepancies as possible have been resolved. For example, Garrison et al. (2006) reported that coders in their study initially achieved 43 percent intercoder reliability but using the negotiated agreement method raised it to 80 percent. Strictly speaking, ascertaining the level of intercoder agreement is not the same as ascertaining the level of intercoder reliability, which is a measure of the

degree to which the coding of two or more equally qualified coders match when everyone's coding is done in isolation from the rest without negotiation. But the negotiated agreement approach is still advantageous in exploratory research like ours where generating new insights is the primary concern (Morrissey 1974:214-15). More important, it is also useful in situations where coding requires great sensitivity not only to obvious meanings but also more subtle meanings, and where coders have different levels of knowledge in this regard. After all, it is possible that discrepancies in coding and therefore suboptimal levels of intercoder reliability occur because some coders are significantly more knowledgeable than others rather than because there are inherent problems with the coding scheme per se.

With this in mind, the PI unitized and coded another full-length transcript selected at random. Next his research assistant coded the unitized but uncoded copy. We reviewed and adjusted the coded transcripts using the negotiated agreement method and then calculated intercoder agreement. In each discrepant case, we kept track of whether we achieved reconciliation and if so which way it went—that is, whether the PI deferred to the research assistant or vice versa. Occasionally, there were irreconcilable differences. As noted above, before negotiating discrepancies, we had achieved 54 percent intercoder reliability for the primary and secondary codes combined and 65 percent intercoder reliability for the primary codes alone. After negotiating discrepancies, we reached 96 percent intercoder agreement for the primary and secondary codes combined and 97 percent for the primary codes alone. We reconciled 91 percent of our initial disagreements. And of these, the research assistant deferred to the PI 69 percent of the time and the PI deferred to the assistant 31 percent of the time. Thus, the PI's coding was more accurate than his assistant's coding in the sense that after negotiation his coding was correct 87 percent of the time and his assistant's was correct 75 percent of the time.

The potential difficulty with this approach, of course, is the interpersonal dynamics that may be involved in negotiation—especially when a less knowledgeable research assistant (e.g., a student) is working with a more knowledgeable PI (e.g., a professor). For example, one can never be entirely sure that the research assistant's willingness to defer to the PI is not due to his respect for the PI's knowledge of the subject matter, intimidation, or something else. Conversely, the assistant's insistence on dissent may stem from a lack of certainty on the PI's part, which could be signaled by something as small as a facial expression, change in posture, or shift in vocal intonation. Or there could be an implicit threshold of dissent where the assistant feels that he should only dissent a few times or not in particularly forceful ways.

Put differently, the unevenly distributed resolutions could be a function of either differential coding accuracy or negotiation dynamics. It is probably impossible to know for sure. But researchers should be aware of this possible dilemma and confront it openly with their assistants.

We discussed why the assistant tended to defer to the PI and agreed that interpersonal dynamics were not a serious problem during coding negotiations. Although the PI was more knowledgeable about the project than his assistant and, therefore, had a better idea of what sort of information needed to be identified and retrieved from the interview transcripts, this did not apparently cause the assistant to defer unnecessarily to the PI. Nor did it prevent the PI from deferring to his assistant. Furthermore, we had been working together for several months by this time and had established a comfortable rapport further reducing the possibility that the assistant was succumbing unwittingly to the PI's authority or cues. Importantly, all of this was consistent with the fact that the PI deferred to his assistant's coding nearly a third of the time and that 9 percent of their differences were irreconcilable. This suggested in particular that the assistant was more than willing to stick to his initial coding decisions when he believed after negotiation that he was correct.

At this point, we felt confident enough in the coding scheme to move from its further development to deployment—that is, using it for coding all 101 transcripts from scratch. The development stages require at least two coders, but the deployment stage does not. We believed that if two or more knowledgeable coders could code the transcripts, then the level of intercoder reliability would be high. Unfortunately, this was not possible in our situation. So the PI coded all the transcripts with the codes requiring deeper knowledge of the subject matter, such as the secondary codes in online appendix (which can be found at http://smr.sagepub.com/supplemental/), because the negotiated agreement procedure showed that for these codes his coding was more accurate than his assistant's coding. His assistant coded all the transcripts with the most straightforward codes, such as names of organizations. Others have argued that during the deployment stage, high values of intercoder reliability justify the choice for one rather than two or more coders using one set of codes (Burla et al. 2008:116; Dunn 1989:37). It follows that high values of intercoder agreement justify the choice of a single coder too as long as the final coder is the one whose coding generally carried the day during the negotiation process. Put differently, by utilizing the negotiated agreement method outlined here a knowledgeable coder can be reasonably confident that his coding would be largely consistent with that of other equally knowledgeable coders if they were available.[7]

Of course, reverting to the intercoder agreement method is unnecessary if an acceptable level of intercoder reliability is achieved in the first place. However, we suspect that with in-depth semistructured interviews and with coders who possess significantly different levels of knowledge about the project's subject matter it will be necessary to use intercoder agreement.

Several additional remarks are important. First, as mentioned earlier throughout the entire process we kept a detailed codebook, which included for each code the code name, the decision rule for what to include or exclude, and an example where necessary. This was refined as we continued to develop the coding scheme and establish intercoder reliability and agreement. This is a very important tool for enhancing reliability, as coding proceeds and enables other users of the data to understand where relevant material may be located in the transcripts.

Second, we agree with others who have argued that all else being equal, simpler coding schemes are better than complex ones. They tend to improve intercoder reliability and agreement, save time, and avoid codes that may turn out later not to be useful. As noted earlier, we took steps to simplify our coding scheme although we still ended up with one that was rather complex. So, we took an additional step to deal with its complexity. We soon discovered that it was good *not* to use the entire code scheme on a single pass through the transcript. There were simply too many codes to keep in mind all at once. Hence, we split the coding scheme into two sets of code families roughly equal in size. Then, we coded each unitized transcript twice—that is, a transcript for which the PI had already identified and bracketed units of text to be coded—first using the first set of code families and then again using the second set. This made coding cognitively easier and improved intercoder reliability and agreement.[8] As we have said, minimizing the number of codes wherever possible also helps keep coders sharp.[9] We also used this process of coding each transcript twice when we deployed our final coding scheme on all 101 transcripts.

Finally, after the coding was finished and data analysis and writing began, the PI still spent considerable time going back to the transcripts to review what people had said. Indeed, doing this kind of research—even with the assistance of coding software—requires that the analyst immerse herself deeply and repeatedly in the transcript data. Coding per se is no substitute for such immersion.

Calculating Reliability and Agreement

Following Miles and Huberman (1984:63) to determine the level of intercoder reliability for a code, we divided the number of times that all coders used it in the same text unit by the number of times that any coder used it in the

transcript. That is, we divided the number of coding agreements by the number of agreements and disagreements combined. For instance, with two coders, if 20 text units had been coded "change" by at least one of them and in 15 of those cases both had invoked the code on the same text unit, then the level of intercoder reliability would be 75 percent (15/20 = .75) for the "change" code. Using this same method, we calculated overall intercoder reliability for all codes as a set by dividing the total number of agreements for all codes by the total number of agreements and disagreements for all codes combined. So, if there were 2 coders and 200 instances when at least one of them invoked a code on a text unit and of these there were 50 instances when both coders did so for the same code on the same unit, then the overall level of intercoder reliability would have been 25 percent (50/200 = .25). We used the same procedure to calculate intercoder agreement based on the number of coding agreements achieved after coders tried to reconcile their differences.

It is not ordinarily recommended that intercoder reliability be calculated simply as the percentage of agreement among coders—the so-called *proportion agreement* method (Morrissey 1974)—because this does not take into consideration the possibility that coders might agree occasionally by chance (Bernard 2000:459-61). Chance may inflate agreement percentages, especially with only two coders and when they have only a few codes (Grayson and Rust 2001). Hence, a variety of more complicated statistics are available for calculating intercoder reliability and by extension intercoder agreement (e.g., Armstrong et al. 1997; Grayson and Rust 2001). One of the most common is Krippendorff's α coefficient—a complicated statistic utilized to determine the degree to which coders coding the same text are doing so in a reliable manner after correcting for the possibility that coders may agree by chance (Krippendorff 2004:chap. 11). This was inappropriate in our case. First, the use of α is based on the assumption that all codes have equal probability of being used. This assumption did not hold in our situation. For instance, when discussing how organizations changed, some may have changed in response to competitive pressures but others might not. Hence, the secondary code "cause-competition" might not have appeared in every transcript, whereas the secondary code "research-methods/data" appeared in all transcripts because everyone discussed their organization's research methods. Second, the use of α assumes that all coders have the same qualifications. This criterion was not fulfilled because the PI was more knowledgeable about the subject matter being coded than his assistant—a condition for many projects using in-depth semistructured interview data.

There were other reasons as well why we used the simple proportion agreement method rather than a more complex statistic. First, we had a large number

of codes, which reduced the likelihood that coders agreed by chance (Grayson and Rust 2001). Second, the possibility for multiple codings on a text unit, which was characteristic of our coding, creates problems for calculating intercoder reliability (or intercoder agreement) with some statistics because they require that only one code is applied to a unit of text (Burla et al. 2008). Third, our intent in coding in the first place was not to generate variables for use in statistical analysis as is often the case when complex intercoder reliability statistics are used. When it comes to qualitative analysis in which the goal is the systematic and rule-guided classification and retrieval of text the use of such statistics is less imperative. Finally, ours was an exploratory study for which other researchers have argued that the simple proportion agreement method is an acceptable approach (Kurasaki 2000). That said, researchers should use whatever statistic is appropriate for their situation.

To return to an issue raised earlier, there is, unfortunately, no agreed upon threshold for what constitutes a numerically satisfactory level of agreement among coders. Nevertheless, the literature does provide some guidance. Our initial aim was to achieve a discriminant coding capability of 80 percent to 90 percent (Miles and Huberman 1984:63). As noted above, our intercoder agreement scores surpassed this but not our intercoder reliability scores. It is worth noting, however, that what passes for an acceptable level of intercoder reliability varies considerably in the literature according to the standards of different researchers as well as the method of calculation (Dunn 1989:37). For instance, Hodson (1999:51) says that an "inter-coder correlation" of 79 percent ensures a "relatively high degree of reliability." Fahy (2001) held that an intercoder reliability range of 70 percent to 94 percent was "acceptable" to "exceptional" for his analysis of transcripts from conference discussions. Kurasaki (2000) reported intercoder agreement scores of 70 percent during coder training and 94 percent during actual coding—both of which were deemed acceptable. Others argue that if the research is exploratory, looser standards are permissible (Hruschka et al. 2004:313; Krippendorff 2004:chap. 11). And some say that in making decisions about cutoff points, consideration should be given to the range and size of the coding scheme and that if the study is concerned with only a "rough estimate" of a code's population prevalence then a looser standard is acceptable as well (Hruschka et al. 2004:326). Krippendorff (2004:241-43; see also Bernard 2000:461) argues that there is no set answer for how reliable is reliable enough. He maintains that standards for data reliability should not be adopted ad hoc, but must be related to validity requirements imposed upon research results, specifically to the costs of drawing wrong conclusions, which, for example, could be more serious in medical as opposed to sociological research. He also says that, "If it is an exploratory

study without serious consequences … [the] level may be relaxed considerably, but it should not be so low that the findings can no longer be taken seriously."

Some Alternative Approaches

Despite the fact that there is very little literature on the subject, ours is not the only conceivable way in which researchers might develop reliable coding schemes for in-depth semistructured interviews. Depending on the circumstances, such as the nature of the interview questions, the complexity of the coding scheme, and the number and knowledge levels of the coders available, other approaches may be on offer.[10] To investigate this at the suggestion of one anonymous reviewer, we contacted researchers on seven other projects that involved this type of data to find out how they developed their own reliable coding schemes. The projects focused on various topics including the management of medical clinics, people's personal privacy issues, the behavior of welfare recipients, compliance with federal regulations, sexual identity, sexuality, and the experience of living in urban poverty. The number of transcripts involved in these projects ranged from about 30 to over 200. We found that approaches were wide ranging and depended somewhat on how much financial support researchers had for hiring coders.

One researcher took an entirely inductive approach to coding without any pregiven coding scheme and entirely ignored the question of coding reliability. This researcher worked alone, immersed herself in the transcripts, searched for common themes across transcripts, and created word processing files that she called *buckets* (i.e., codes) for each theme that emerged inductively. Coding involved using a word processing package to literally cut and paste relevant portions of transcripts into the appropriate buckets. She did not check for coding reliability believing that what mattered was simply conveying to the reader her interpretation of the data. Another researcher we talked with mentioned that she was aware of other projects that did essentially the same thing.

A few other researchers told us that they also used an inductive approach to coding. Others, however, began instead with a pregiven set of codes as we did. In any case, these projects all involved research assistants. In some cases, researchers and assistants each coded *all* of the transcripts. Later, they compared the two sets of coding, discussed their coding disagreements, and tried to resolve them as best they could. Then, they calculated intercoder agreement. In other cases, researchers and assistants coded only a small *subset* of transcripts, compared coding, discussed, and resolved disagreements as best they

could, calculated intercoder agreement, and then adjusted the code scheme as necessary before coding all the transcripts. In several cases, researchers reported running into the same unitization problem that we did and liked our approach for handling the problem once we described it to them.

For one large, well-funded project that involved several coders the project manager told us that they began with a pregiven set of codes. Then they trained coders to do a "first layer" of transcript coding on a transcript. This involved only the broadest thematic codes (i.e., primary codes) and using them to code large chunks of a transcript. Then they checked for intercoder reliability and repeated the exercise until the coders achieved a satisfactory outcome of 90 percent reliability or better. The intercoder reliability check involved comparing the assistants' coding to a "master coding" performed by the PI who presumably knew best what to look for in the transcripts. In other words, the PI established a gold standard against which other coders were compared. Discrepancies in coding were discussed, coding instructions were clarified, but the codes themselves were not changed. After the fourth iteration, the coders who still failed to code reliably enough were dismissed from the team. Once reliable coders were in place, each was assigned a subset of codes in which to specialize and coding proceeded—including occasionally adding new codes inductively, which they called *second layer* codes (i.e., secondary codes). Again, however, in this project the unitization problem surfaced and the team found it tricky to handle. They agreed eventually to use the individual printed transcript page as the unit of analysis. They were coding by hand rather than using electronic software and computers as the rest of the projects we queried did as well as our own. The project manager lamented the fact that there was virtually no literature to guide them as they developed this approach through a largely trial-and-error process.

We learned as well that in some cases like the previous one, once reliability for coding the primary codes was established in this manner on a subset of transcripts, assistants then coded large thematic chunks of all the transcripts using the primary codes. Then the PI working on a particular theme returned to these coded transcripts and coded the relevant chunks again but with a more detailed set of secondary codes, which were often generated more inductively than deductively. This coding required a high degree of knowledge, which is why the PI or sometimes a team of very experienced coders did it.[11] One would, of course, still want to check how reliable this secondary coding was.

Four things are important here. First, some projects used pregiven codes; some projects used inductively generated codes; and some projects used both like we did. Second, some projects ignored the question of intercoder

reliability and agreement entirely but most did not. Third, those projects that tried to establish intercoder reliability or agreement ran into the unitization problem as we did but most were not entirely comfortable with how they had tried to resolve it, which is why they appreciated learning how we handled it systematically. Most handled it in an ad hoc fashion simply trying to determine as best they could on a case-by-case basis whether two or more coders had coded the same way in a particular instance even though they had unitized a portion of text somewhat differently. Finally, several researchers wished that they had had more guidance from the literature on how to handle these issues and so were generally pleased to learn of our article. In sum, a variety of approaches for developing reliable coding schemes for in-depth semistructured interview transcripts are on offer depending on the resources available to researchers. We hope that our approach will help researchers working with such data—particularly researchers without large budgets for hiring a team of coders—and regardless of whether they use an inductive, deductive, or mixed approach to code development.

Discussion and Conclusion

No method is perfect. As noted earlier, one potential limitation of the methods reported here involves unitization. The way transcripts are unitized may bias coding and inflate intercoder reliability (or agreement) scores. We explained why we doubted that this was a problem for us. However, further research should be done to investigate the relationship between unitization and intercoder reliability and agreement to see if indeed this is a problem. For example, one way to explore this would be to unitize the same transcript in two ways. The first version would bracket only small portions of text that spoke directly to a particular coded issue, while the second version would bracket those portions plus whatever surrounding text provided contextual or background information. Then the same two people could code each version, intercoder reliability for each version could be calculated, and then the two intercoder reliability scores could be compared to see if there were significant differences in how each version was coded. If there was no significant difference, then we would be less concerned that unitization would bias intercoder reliability.

The approach we have outlined is likely useful in slightly modified form under a wider range of circumstances than those described earlier. First, some studies using interview transcripts have multiple PIs. If so, then there is probably more than one knowledgeable coder available in which case additional people may be incorporated into the unitization and negotiated

agreement processes. Moreover, the reproducibility of the unitizing process could be checked with the same sort of approach that we used to establish intercoder reliability and agreement. That is, two PIs could each unitize the same transcript, compare their unitizations, and if necessary negotiate their differences before using that transcript to work on the coding scheme.

Second, sometimes the knowledgeable PI does not have time to do the coding and needs to hire a team of research assistants to do it all. What then? The PI can still participate in developing the coding scheme and establishing intercoder reliability and agreement along the lines we have suggested. However, during the process, she can spend enough time educating the research assistants in what to look for when coding that they become knowledgeable enough to do the ultimate coding themselves. She could also use the negotiated agreement process to see if some coders are better at coding with certain codes than others. In both cases, this might eventually lead the PI to establish a division of labor among coders by assigning a few specific code families to each assistant thereby making it easier for them to become knowledgeable experts with respect to their particular subset of codes. This is reminiscent of one of the alternative approaches we reviewed earlier. The limit here is how knowledgeable the coder must be in the subject at hand. This depends in part on how specific and narrowly defined the interview questions are and how much nuance is required in interpreting and coding the answers. The more open-ended and general the questions and answers are and the more vague the codes are, the more difficult it would be to educate the research assistants to a point where there was sufficient confidence in their coding.

Third, qualitative data analysis software including Atlas.ti is available nowadays that can handle documents, pictures, and even digital and video recordings as well as interview transcripts. As a result, the basic procedures that we have offered for unitizing qualitative data and establishing the discriminant capability of coding schemes can be used in research based on all sorts of qualitative data. For example, this method is appropriate for those working with long text documents like government reports, histories, or ethnographies that need complex coding schemes and that require texts to be unitized for coding. It can also be used for coding video tape and images. Photographs or art work in pdf format can be unitized and code schemes for them can be developed according to our method. In turn, the level of intercoder reliability and agreement can be calculated as well.

There are, however, situations for which our method may be less suitable. Notably, some qualitative researchers might object that our approach is based on a set of codes derived from initial hypotheses and insights prior to coding

rather than a less positivist approach based on grounded theory where the codes emerge inductively as themes are identified during coding (Glaser and Strauss 1967). Assessing intercoder reliability or agreement under these conditions is very tricky and little work has been done to figure out how to do it (but see, Armstrong et al. 1997; Popping 1992). This is why intercoder reliability and agreement checks are usually performed for text analysis only when a preconceived coding scheme with a fixed number of codes is being used—not when the grounded approach is being used (Hodson 1999:chap. 3; Krippendorff 2004:chap. 11). That said, once inductive codes have been developed from the analysis of a subset of transcripts, they can then be deployed like any other code and with the same precision as suggested by our approach for dealing with intercoder reliability and agreement.

Deductive and inductive approaches are not necessarily mutually exclusive. It is possible as coding proceeds during the deployment stage that a coder working with a fixed set of codes developed with our method may still identify and therefore need to code new themes that were not anticipated previously. This is common when analyzing in-depth semistructured interviews especially when studying phenomena for which there is little previous research (e.g., Miles and Huberman 1984:56-60; Rubin and Rubin 2005). One way to handle this would be to add new codes to the coding scheme but this risks compromising the levels of intercoder reliability and intercoder agreement already achieved. Another approach is to attach a series of separate "memos" to various pieces of text—an option available in Atlas.ti and presumably other qualitative data analysis software. As with the software's coding function, the memo function enables a coder to bracket a portion of text where an important theme or concept appears and then write an explanatory memo for it. Through this so-called *free coding approach*, one can gradually build up what amounts in effect to a second more grounded coding scheme in addition to the first one. Thus, the original set of codes for which intercoder reliability and agreement are established remains unchanged while as coding proceeds an independent set of memos is created inductively and deployed. Researchers have advocated mixed-method approaches to social science (Tashakkori and Teddlie 1998). It follows that combining these two approaches may be more fruitful than using either one alone.

To conclude, by bringing together processes for unitizing transcripts and establishing intercoder reliability and agreement, this article has offered an innovative method for developing coding schemes for use with in-depth semistructured interview data. It includes techniques for establishing intercoder reliability and intercoder agreement that are especially—but not exclusively—appropriate for situations where coding of such data requires a high

level of knowledge of the interview subject matter but where only one knowledgeable coder is available. This method helps rectify a surprising omission in the qualitative methods literature. And it is a method that is eminently practical, given the advent of qualitative data analysis software. With this in mind, we hope that researchers using this type of data will be more explicit in how they approach the issue of coding reliability.

Acknowledgments

Thanks for comments on the previous versions of this article go to Denise Anthony, Marc Dixon, Kathryn Lively, Janice McCabe, Kathleen Sherrieb, and two anonymous reviewers.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Support for this project was provided by a grant from the U.S. National Science Foundation (SES 0813633), the Rockefeller Center for Public Policy and the Social Sciences at Dartmouth College, and the Sonic Project at the Copenhagen Business School.

Notes

1. We checked all the back issues of *Sociological Methods and Research* and found no articles on this topic. We also contacted two editors (past and current) of two of the top sociological methods journals who said they were unaware of any such articles.
2. We contacted two prominent research teams using this type of data to see if they at least had documentation about how they established coding reliability. They did not.
3. Developing an acceptable level of intercoder reliability or agreement often takes considerable time. For example, Hruschka et al. (2004) reported that one study based on 24 in-depth semistructured interviews with responses to each question running in length up to one page of text took eight rounds of coding in order to establish an acceptable intercoder reliability level. We took 11 rounds.
4. We determined early in this process that intercoder reliability was very high for coding organization names—typically over 80 percent reliability. So in order not to inflate our reliability scores, we excluded organization name codes from this process. For the same reason, others have dropped such nonsubstantive codes from calculations of intercoder reliability (e.g., Burla et al. 2008:115).

5. A related problem was that some transcripts were easier to code than others. This is something that is barely mentioned in the literature (but see, Hruschka et al. 2004). It depended on how clear the respondent's answers to our questions were and whether the respondent spoke coherently. Hence, one should test a coding scheme's reliability (or agreement) on more than one transcript before accepting it as final and moving on to deployment.
6. These are sometimes referred to as the *surface* and *latent* meanings of a text, respectively (Popping 2010:1068).
7. Of course, discrepancies in intercoder agreement may also arise when different interpretations based on different theoretical priors are applied to the same text. In other words, it is not always a matter of differences in knowledge per se that might account for coding discrepancies. This is all the more reason for providing coders with adequate training—including familiarity with the theoretical framework informing the project.
8. The principal investigator (PI) came to suspect later after the ultimate coding had begun that doing the first pass through a transcript in the afternoon and then the second pass the following morning enabled him to remain sharper and more focused mentally than if he had done both codings of the transcript on the same day.
9. To make coding cognitively easier, some practitioners divide the transcript material into several topical areas even before coding begins. Then they code each area of material separately. This would not have been feasible in our case because our transcripts were too free flowing with many passages of text in which more than one topical area was discussed at the same time.
10. Krippendorff (1995, 2004), for example, has detailed discussions of other options and the conditions under which they are appropriate.
11. We thank an anonymous reviewer for this insight.

References

Armstrong, David, Ann Gosling, Josh Weinman, and Theresa Martaeu. 1997. "The Place of Inter-rater Reliability in Qualitative Research: An Empirical Study." *Sociology* 31:597-607.

Becker, Howard. 1970. "Problems of Inference in Participant Observation." Pp. 189-201 in *Qualitative Methodology*, edited by William Filstead. Chicago, IL: Rand McNally.

Bernard, H. Russell. 2000. *Social Research Methods: Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage.

Burla, Laila, Birte Knierim, Jürgen Barth, Katharina Liewald, Margreet Duetz, and Thomas Abel. 2008. "From Text to Codings: Inter-coder Reliability Assessment in Qualitative Content Analysis." *Nursing Research* 57:113-17.

Calarco, Jessica. 2011. "'I Need Help!' Social Class and Children's Help-seeking in Elementary School." *American Sociological Review* 76:862-82.

Campbell, John L. and Ove K. Pedersen. 2014. *The National Origins of Policy Ideas: Knowledge Regimes in the United States, France, Germany and Denmark*. Princeton, NJ: Princeton University Press.

DeSoucey, Michaela. 2010. "Gastronationalism: Food Traditions and Authenticity Politics in the European Union." *American Sociological Review* 75:432-55.

Deutscher, Irwin. 1970. "Looking Backward: Case Studies on the Progress of Methodology in Sociological Research." Pp. 202-16 in *Qualitative Methodology*, edited by William Filstead. Chicago, IL: Rand McNally.

Dunn, Graham. 1989. *Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors*. New York: Oxford University Press.

Edin, Kathryn and Maria Kefalas. 2005. *Promises I Can Keep: Why Poor Women Put Motherhood Before Marriage*. Berkeley, CA: University of California Press.

Edin, Kathryn and Laura Lein. 1997. *Making Ends Meet: How Single Mothers Survive Welfare and Low-wage Work*. New York: Russell Sage Foundation.

Engel, Mimi. 2007. "Mixing Methods: Reliability and Validity across Quantitative and Qualitative Measures of Relationship Quality." Pp. 255-76 in *Unmarried Couples with Children*, edited by Paula England and Kathryn Edin. New York: Russell Sage Foundation.

England, Paula and Kathryn Edin. 2007. "Unmarried Couples with Children: Hoping for Love and the White Picket Fence." Pp. 3-21 in *Unmarried Couples with Children*, edited by Paula England and Kathryn Edin. New York: Russell Sage Foundation.

Fahy, Patrick. 2001. "Addressing Some Common Problems in Transcript Analysis." *The International Review of Research in Open and Distance Learning* 1. (http://www.irrodl.org/index/php/irrodl/article/view/321).

Garrison, D. R., M. Cleveland-Innes, Marguerite Koole, and James Kappelman. 2006. "Revisiting Methodological Issues in Transcript Analysis: Negotiated Coding and Reliability." *Internet and Higher Education* 9:1-8.

Gerson, Kathleen. 1985. *Hard Choices*. Berkeley: University of California Press.

Glaser, Barney and Anselm Strauss. 1967. *The Discovery of Grounded Theory*. New York: Aldine de Gruyter.

Grayson, Kent and Roland Rust. 2001. "Interrater Reliability Assessment in Content Analysis." *Journal of Consumer Psychology* 10:71-73.

Greene, Jennifer C., Valeri J. Caracelli, and Wendy R. Graham. 1989. "Toward a Conceptual Framework for Mixed-method Evaluation Designs." *Educational Evaluation and Policy Analysis* 11:255-74.

Halliday, Terence C. and Bruce G. Carruthers. 2009. *Bankrupt: Global Lawmaking and Systemic Financial Crisis*. Stanford, CA: Stanford University Press.

Hodson, Randy. 1999. *Analyzing Documentary Accounts*. Thousand Oaks, CA: Sage.

Hruschka, Daniel, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard Jenkins, and James Carey. 2004. "Reliability in Coding Open-ended Data: Lessons Learned from HIV Behavioral Research." *Field Methods* 16:307-31.

Krippendorff, Klaus. 1995. "On the Reliability of Unitizing Continuous Data." *Sociological Methodology* 25:47-76.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Thousand Oaks, CA: Sage.

Kurasaki, Karen S. 2000. "Intercoder Reliability from Validating Conclusions Drawn from Open-ended Interview Data." *Field Methods* 12:179-94.

Lewins, Ann and Christina Silver. 2007. *Using Software in Qualitative Research*. Los Angeles, CA: Sage.

Lombard, Matthew, Jennifer Snyder-Duch, and Cheryl Bracken. 2002. "Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability." *Human Communication Research* 28:587-604.

Mason, Jennifer. 2002. *Qualitative Researching*. 2nd ed. Thousand Oaks, CA: Sage.

McCabe, Janice. 2011. "Doing Multiculturalism: An Interactionist Analysis of the Practices of a Multicultural Sorority." *Journal of Contemporary Ethnography* 40:521-59.

Miles, Matthew B. and A. Michael Huberman. 1984. *Qualitative Data Analysis: A Sourcebook of New Methods*. Beverly Hills, CA: Sage.

Morrissey, Elizabeth R. 1974. "Sources of Error in the Coding of Questionnaire Data." *Sociological Methods and Research* 3:209-32.

Popping, Roel. 1992. "In Search of One Set of Categories." *Quality and Quantity* 26: 147-55.

Popping, Roel. 2010. "Some Views on Agreement To Be Used in Content Analysis." *Quality and Quantity* 44:1067-78.

Riffe, Daniel and Alan Freitag. 1997. "A Content Analysis of Content Analysis: Twenty-five Years of Journalism Quarterly." *Journalism and Mass Communications Quarterly* 74:873-82.

Rosenthal, Robert. 1987. *Judgment Studies: Design, Analysis, and Meta-Analysis*. New York: Cambridge University Press.

Rubin, Herbert J. and Irene S. Rubin. 2005. *Qualitative Interviewing: The Art of Hearing Data*. 2nd ed. Thousand Oaks, CA: Sage.

Strang, David. 2010. *Learning by Example: Imitation and Innovation at a Global Bank*. Princeton, NJ: Princeton University Press.

Tashakkori, Abbas and Charles Teddlie. 1998. *Mixed Methodology: Combining Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage.

Useem, Michael. 1984. *The Inner Circle: Large Corporations and the Rise of Business Political Activity in the U.S. and U.K*. New York: Oxford University Press.

Weber, Robert Philip. 1990. *Basic Content Analysis*. 2nd ed. Newbury Park, CA: Sage.

Wilson, William Julius. 1996. *When Work Disappears: The World of the New Urban Poor*. New York: Alfred A. Knopf.

Zelditch, Morris. 1970. "Some Methodological Problems of Field Studies." Pp. 217-34 in *Qualitative Methodology,* edited by William Filstead. Chicago, IL: Rand McNally.

Author Biographies

John L. Campbell is the Class of 1925 professor in the Department of Sociology, Dartmouth College, USA, and professor of political economy in the Department of Business and Politics, Copenhagen Business School, Denmark. His current work includes a study of how expert policy ideas are developed and deployed in advanced industrial economies, which is entitled *The National Origins of Policy Ideas: Knowledge Regimes in the United States, France, Germany and Denmark* (Princeton University Press, forthcoming, coauthored with Ove Pedersen), and another study of why small nation-states tend to perform so well in today's global economy.

Charles Quincy is a recent graduate of Dartmouth College. He is currently an industry solutions consultant at the IBM Corporation in New York City.

Jordan Osserman is a recent graduate of Dartmouth College. He is currently a graduate student in the psychoanalysis, history, and culture program at Birkbeck College, University of London.

Ove K. Pedersen is professor of comparative political economy, Department of Business and Politics, Copenhagen Business School, Denmark. In addition to his research on expert policy ideas (with John Campbell), his most recent work includes the analysis of Danish institutional competitiveness published as *Konkurrencestaten* (Hans Reitzels Forlag, 2011) and a collaborative study of the influence of intellectuals on nation building, which is soon to be published as *The Role of Legends in Nation Building* (McGill-Queens University Press and DJØF Publisher, forthcoming).