Beyond a Coefficient: An Interactive Process for Achieving Inter-rater Consistency in Qualitative Coding

Vonna L Hemmler

Allison W Kenney

Susan Dulong Langley

Carolyn M Callahan

E. Jean Gubbins

Shannon Holder

Hemmler, V. L., Kenney, A. W., Langley, S. D., Callahan, C. M., Gubbins, E. J., & Holder, S. (2022). Beyond a coefficient: an interactive process for achieving inter-rater consistency in qualitative coding. *Qualitative Research, 22*(2), 194-219. https://doi.org/10.1177/1468794120976072

THIS WORK WAS SUPPORTED BY U.S. DEPARTMENT OF EDUCATION INSTITUTE OF EDUCATION SCIENCES THROUGH PR/AWARD #R305C140018. THE OPINIONS EXPRESSED ARE THOSE OF THE AUTHORS AND DO NOT REPRESENT VIEWS OF THE INSTITUTE OF EDUCATION SCIENCES OR THE U.S. DEPARTMENT OF EDUCATION.

Abstract

Though qualitative research has become more prevalent in practice over the last 30 years, there is still considerable uncertainty among researchers regarding how to ensure inter-rater consistency when teams are tasked with coding qualitative data. In this article, we offer an explanation of a methodology our qualitative team used to achieve systematic coding of our dataset in a way that preserved the contextual, subjective nature of the data, lent itself to the deductive and inductive creation of a layered codebook, and ensured consistent application of the codebook to varied types of data. This methodology prepared us to draw logical and substantiated conclusions during subsequent analyses; hence, the process serves as a welcome addition to the literature on consistently coding qualitative data in a manner that honors its defining characteristics.

Keywords: qualitative research, qualitative coding, inter-rater reliability/consistency, consensus coding, team coding, inductive and deductive coding

Qualitative research has increased in frequency of application and prominence across scholarly fields over the past 30 years (Elliott et al., 1999; Elmore and Woehlke, 1998; Lombard et al., 2002; Mays and Pope, 2000; Smith, 1987), so much so that organizations as of late have published standards for reporting and reviewing qualitative work in journal publications (e.g., Levitt et al., 2018). Despite this increase in acceptance, dialogue about qualitative inquiry continues. Some tout its dynamic, reflexive, open-ended, contextually-sensitive, and interactive nature as qualities that equip qualitative methodologists to answer questions that more positivist, quantitative methods cannot (e.g., Levitt et al., 2018; Silverman, 2000; Tobin and Begley, 2004). Others cite these same characteristics as reasons for qualitative research's perceived unreliability and lack of rigor (e.g., Carey et al., 1996; Tobin and Begley, 2004; Weston et al., 2001).

We position ourselves in alignment with the former group, as we see value in the "rich, detailed, and heavily contextualized" nature of qualitative data and accept it as "legitimate data for analyses" (Fereday and Muir-Cochrane, 2006; Levitt et al., 2018: 27). We similarly value the emphasis in the qualitative tradition on inductive, open-ended discovery and recursive reflection (Levitt et al., 2018; Silverman, 2000). However, we also recognize that it is precisely these characteristics that have caused detractors of qualitative research to argue against this mode of inquiry, particularly when it comes to reliability (Sandelowski, 1986: 33). We understand why some qualitative researchers have problematized a traditional approach to reliability, and in response to this critique, we offer a new perspective within the alternative approach that has viewed reliability within qualitative research in terms of consistency and transparency of coding methods.

To this end, we describe the iterative, consensus processes (Hill et al., 1997, 2005; Levitt et al., 2018) four coders used to achieve inter-rater consistency in coding qualitative data collected as part of a study of an elementary gifted education program in a US public school district. Although we frame our methods in this study, we believe they are applicable to any qualitative inquiry across the social sciences that seeks to utilize a team-based approach to make sense of varied types of complex data. Our methods are innovative in their contemporaneous combination of several established practices.

First, we simultaneously employed deductive and inductive coding schemes in our thematic analysis (Braun and Clarke, 2006). Second, our data were collected via structured, semi-structured, and unstructured means, and our literature search did not identify any studies in which existing coding schemes were applied to varied types of data in this manner. This complex nature of our data in turn necessitated a multilayered codebook, the accurate application of which would be difficult to assess with a traditional measure of reliability. Finally, we applied the previous three features within a team-based, consensus approach to coding (Hill et al., 1997, 2005). Thus, we innovatively interfaced these practices in a way that created a new approach to team-coding qualitative data that serves the same purpose as reliability because of its consistent and transparent nature.

Our detailed explanation of our processes incidentally answers the calls in the literature a) for qualitative researchers to offer concrete evidence of their methodologies (Aguinis et al., 2019; Glazer and Egan, 2018; Kirk and Miller, 1986; Peterson, 2019; Stearns et al., 2019), and b) for qualitative reports to be evaluated "in terms of their own logic of inquiry" (Levitt et al., 2018: 27-28; Syed and Nelson, 2015). We both see the need for and contribute to this "logic of inquiry" by offering a new approach to qualitative coding that yields trustworthy results, thus foregrounding the product of the research over complete control of the process (Sandelowski, 1986; Syed and Nelson, 2015). In laying bare the steps of our active, negotiated, and social process to show how we used them to consistently reach concordance among coders, we have conveyed a way of achieving a sense of coding consistency in lieu of a more traditional measure of it. Thus, our methods contribute to the "trustworthiness of the method" of qualitative inquiry (Eisner, 1998; Garrison et al., 2006; Hill et al., 1997: 517; Hruschka et al., 2004; Kurasaki, 2000) by serving the same purpose as a more conventionally-reached coefficient of inter-rater reliability (IRR): to assure readers and reviewers of qualitative work that conclusions reached as a result of our rigorous processes are scientifically sound. Scholars have identified "increasingly sophisticated procedures to guide the interpretive acts of social researchers" (Kirk and Miller, 1986: 5) in this way, and we contend the coding methodology articulated here is one such procedure.

Qualitative Coding and Reliability

Saldaña (2016: 4) defined a qualitative *code* as "a word or short phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute for a portion of . . . data." He and others have proposed, then, that qualitative coding is a "heuristic" (Saldaña, 2016: 42) between collecting data and interpreting it during analysis (Charmaz, 2006).

While there are numerous ways to engage in qualitative coding, the most common way to verify the soundness of the processes, particularly when coding is team-based, has been through establishing IRR. However, researchers have disagreed over whether IRR as a construct in qualitative work is necessary or sufficient (Armstrong et al., 1997; Carey et al., 1996; Hodson, 1998; Marshall and Rossman, 1989; Morse, 1999; Morse et al., 2002; Saldaña, 2016; Silverman, 2000; Weston et al., 2001). Perhaps confounding this debate is the complicated nature of reliability itself and the options available to establish it (Gwet, 2014; Hruschka et al., 2004).

Reliability generally refers to arriving at consistent answers when repeatedly using the same instrument to measure something (Bernard, 2000). For this article, we focus on the "consistent" element of this definition to view inter-coder agreement as concordance among coders that produces consistent results from analytical procedures over time (Lincoln and Guba, 1985; Noble and Smith, 2015). Early qualitative studies borrowed from quantitative approaches and reported a percentage of agreement among coders for IRR (Carey et al., 1996). Methods for calculating IRR beyond percentage agreement have since emerged to account for chance, the most common of which is Cohen's kappa – though its use does not come without caveats, such as the impact of study sample size, length and complexity of data, and complexity of codebook; the lack of consensus on an appropriate benchmark for the coefficient (Church et al., 2019); and the understanding that any measurement of IRR comes with some imprecision (Burla et al., 2008; Carey et al., 1996; Gwet, 2014; Hayes and Krippendorff, 2007; Hruschka et al., 2004; MacPhail et al., 2016; Miles and Huberman, 1994; Smith and McGannon, 2018; Syed and Nelson, 2015).

Some advocate for abandoning this statistical approach to reliability in interpretivist work in favor of other guidelines, such as trustworthiness as it is demonstrated through credibility, transferability, and dependability (Hammer and Berland, 2014; Hayes and Krippendorff, 2007; Hill et al., 2005; Lincoln and Guba, 1985; Schoenfeld, 1992; Smith and McGannon, 2018). This shift balances the value within qualitative work on varied interpretations with the assurance that researchers do not approach the data solely through their own idiosyncrasies and thus arrive at inaccurate coding agreement (Hill et al., 2005; Hruschka et al., 2004; Smith and McGannon, 2008). If there is momentum in the field of qualitative inquiry toward valuing trustworthiness and consistency in coding over a more traditional understanding of IRR, then one means to this end is to offer complete transparency in the description of the research processes, as we do here.

Rigor

Underlying this debate surrounding IRR is the notion of *rigor*, which has traditionally been associated with the concepts of reliability and validity as they are understood in quantitative approaches (Silverman, 2000). However, some have lamented that these concepts are inappropriately utilized in arenas where they are not relevant, like qualitative inquiry (e.g., Golafshani, 2003; Guba and Lincoln, 1994; Tobin and Begley, 2004; Tracy, 2010). As a result, some scholars have equated rigor not with reliability, but instead with the consistency and transparency, and thus trustworthiness, of a methodology, which are brought about by presentation of processes in a manner that enables reviewers to make fair, believable judgments about the processes and their worth (e.g., Aguinis et al., 2019; Lincoln and Guba, 1985). By outlining our consistent, believable, and fair coding processes in a tightly detailed, cohesive, and transparent manner, we display the rigorous nature of our methods that establish coding consistency in lieu of an IRR quotient.

Data Sources

We believe the practices we established in coding our data are applicable beyond our dataset. However, we also recognize the importance of clearly explaining the structure(s) of the data we coded so readers can appreciate the origins of the practices we describe. Our dataset came from a larger study of elementary gifted programming and included 147 observations of public elementary school classrooms, 104 teacher and administrator interviews, and extensive fieldnote narratives from a total of 15 school site visits. Observational data were collected via a three-part observation form: a structured portion where observers were prompted to look for particular classroom phenomena; a semi-structured portion where observers commented on six theoretically-derived domains; and an unstructured portion where observers were encouraged to document what was occurring in the classroom in as much detail as possible. Teacher and administrator interviews were all semi-structured. Fieldnote narratives were unstructured and included additional details on observations and interviews as well as reflections and comments on the entire school visit.

Methods

In this section we describe the processes used to code our data in a way that was responsive to the deductive and inductive goals of the research, the complex nature of the collected data and the layered codes resulting from it, and the dynamics of a team approach to coding. Understanding that a well-defined coding scheme is the foundation of any sense of coding reliability or consistency (Burla et al., 2008), we also provide insight into how these processes established the consistent nature of our coding and propose guidelines for researchers with similar empirical challenges. The disclosure of our processes as such ultimately communicates the trustworthiness of our methods (Nowell et al., 2017). Figure 1 provides a visual representation of the methods we describe here.

Initial Codebook Development and Application

The data coding processes began with three concurrent approaches to codebook development. We devised a coding scheme that was both deductive and inductive; we adapted the codebook and unitization rules to the unique features of the varied types of data as needed; and we built a process to account for uncertainty in code application and possible future changes to the codebook. The first two steps established agreement on the codebook so the four coders could achieve consistency in coding, and the third step accounted for the possibility of change at the group level.[1]

Deductive and Inductive Codes

To offer a qualitative explanation for quantitative findings from our larger study, the research team specifically documented on field visits information about classroom practices and school-level program implementation – parameters which stemmed from the theoretical framework of that study. In addition to drawing conclusions about these deductively-derived themes (Fereday and Muir-Cochrane, 2006), the research team sought to account for emerging themes in the data that might contextualize the quantitative findings in unanticipated, yet noteworthy, ways (Charmaz, 2006; Silverman, 2000). Team members used familiarity with the data – either from participation in the collection phase or from an earlier attempt at establishing reliability – to brainstorm themes that we knew or expected would be important. We added these categories to our developing codebook along with the deductively-derived codes described above, reorganizing and refining the codebook during a team retreat and pilot process.

As we piloted the application of 12 parent codes and series of child and grandchild codes, we examined ways to combine seemingly-redundant codes and disentangle divergent concepts within single codes. Through this process, code definitions were further tightened, and coders built intersubjective agreement on their applications (Charmaz, 2006).

Structured, Semi-structured, and Unstructured Data. Early in our discussions, we determined our coding scheme needed to suit all three types of data. The semi-structured interview protocol was partially informed by the larger study's theoretical framework and research questions, but its format also allowed for interviewers to adjust questions depending on

[1] In this section we highlight the important first steps of developing the codebook and coding process. However, the many specific decisions made to suit the needs of our dataset are beyond the scope of this article. Authors can provide additional detail on project-specific processes upon request.

participant responses. Thus, although coders could use the interview prompts as clues to possible themes in the participant responses, we agreed that it was appropriate for any relevant codes to be applied to the interviews if their definitions were met, regardless of the protocol prompt. We extended this decision to the unstructured fieldnote narratives as well. Table 1 displays how one interview question prompted coders to apply certain codes to this excerpt, but the participant's response prompted additional code applications.

The classroom observation forms included several sections. The unstructured section prompted observers to take general notes on what they saw in the classroom. On the semi-structured section of the forms, observers documented what they saw according to six theoretically-derived themes. However, where a practice from this list was not observed, there may have only been a note of "not observed." In this case, no existing code was applied, though we viewed these data as valuable and worth capturing, as they signified "non-examples," or instances in which researchers expected to see something in the field but did not (Table 2). We expanded the codebook to include unique codes for each semi-structured and structured observation prompt, allowing us to use our codes as an indexing device that marked data that were or were not there (see Tables 3A-3D; MacQueen et al., 1998; see also Hruschka et al., 2004).

Unitization

The process we used to begin our coding and establish consistency required that coders first agree on the appropriate length of a unit of data to code (MacQueen et al., 1998; Smith and McGannon, 2008). We could only discuss whether the accurate codes were applied to an interview or observation response if all coders looked at the same bounded set of words. (Semi-) structured observation data were separated into discrete units according to each prompt and the accompanying notes. However, the semi-structured interview and unstructured narrative data meant that many themes could be included in a single response, so we had to determine whether to break up the data into very small units by theme, which may strip them of important context, or keep long responses intact even if they comprised many themes and/or a series of follow-up questions and answers.

For interviews, we set a preliminary rule to start a new unit of analysis with every new interview topic. However, when we jointly coded a portion of an interview and found that several were conducted in a conversational format, we amended our rule to mark the bounds of a unit whenever the interviewer moved on to a new protocol question (see Charmaz, 2006). A second attempt to jointly code the sample interview with this new rule garnered near universal agreement on unit length, facilitating the spreadsheet calibration process described below. This rule allowed for some interview units to be quite long and inclusive of many codes if respondents were verbose or touched on many key themes in their answers. A similar solution was reached for coding fieldnote narrative data at the paragraph level, though the different writing styles of each researcher meant that some units were longer and thus potentially inclusive of many codes, while other units were shorter and narrower in scope. Our unitization of data in this way not only functioned to establish consistency in the early stages of coding, but it was also helpful when we pulled codes for subsequent analyses.

Uncertainty in Code Application

The last step before coding was to devise a system to allow the team to work individually but continue to support one another when necessary. To this end, we created a code called Not Sure that coders could apply when assistance was needed. The unsure coder would write a note explaining the nature of their uncertainty and the team would meet regularly to reach consensus on how to rectify these concerns (MacQueen et al., 1998). This system, explained in further detail below, ensured that coders were applying codes consistently as coding progressed and facilitated forward momentum by allowing coders to flag concerns for later discussion rather than hindering the individual's overall coding progress until the team could offer support.

Calibration

Initial Stages of Understanding and Applying Codebook. Before beginning coding, the team calibrated by testing the existing codebook and further developing it as new inductive themes emerged – both of which processes would eventually enable us to code the remainder of the data individually with confidence of consistency (Charmaz, 2006; Hruschka et al., 2004). Our goal was to build confidence that there was minimal variation among coders while also not limiting the possibility for change across coders to a more suitable coding scheme.

To facilitate this first stage of calibration, a team member selected one observation and interview from the dataset, and the four coders individually blindly coded each using Dedoose qualitative coding software. The same team member then created a spreadsheet depicting the code applications of each coder and formatted each code's font to reflect the level of agreement across coders by bolding the font of codes selected by one coder (25%), underlining codes selected by two coders (50%), italicizing codes selected by three coders (75%), and leaving in plain text codes selected by all coders (100%). Table 4 shows an excerpt with mixed levels of agreement that the team co-coded early in the process, with only two of the finalized codes being agreed on by all coders and thus skipped over in discussions. Table 5 presents a later example of a co-coded excerpt in which there was significant improvement in alignment on applied codes. In this instance, six of the finalized codes were agreed upon by all coders and thus not discussed in meetings.

In a series of meetings, the team reviewed the spreadsheet, code by code (aside from those codes that showed 100% agreement, as described above), which allowed each coder to explain their code choices. The instances of 50% agreement in particular typically provided fodder for rich team discussion, as it meant there was enough interpretation by two coders to consider applying the code (see Table 4). In all discussions, all coders were encouraged to share their thought processes without fear of being an outlier until coming to a four-coder consensus in both code application and code definition (see Hill et al., 1997; 2005). Although this process did not require us to discuss every possible code in the codebook (only those that coders had applied in their individual coding), the first observation documents we calibrated did include discussions of at least every classroom-relevant parent code. The first interview transcript likewise required discussions of every classroom- and school-level parent codes. Subsequent calibration on interviews and observations were selected to prioritize discussions of codes that were less commonly applied, which ensured we had touched on every aspect of the codebook during the calibration process.

When considering percentages of code agreement, it is important to note that instances of 25% agreement were not necessarily a case of aberrant code application. Although there were occasions when one coder no longer found merit in applying a particular code (Table 6), there were a number of other times when a coder's experience with or sensitivity to a topic helped identify a code application that the rest of the team missed (Tables 4 and 7). As such, this part of the calibration process contributed to coders' common understanding of codes and their capacity for applying them consistently. We also used the results of this process to expand our definitions with examples and non-examples in the codebook.

Improved Agreement of Code Application. We found in this process that we tended to agree quickly on the parent code themes present in an excerpt, but the more nuanced sub-themes in child codes required additional discussion. To work toward agreement in these discussions, coders first shared their thought processes when reading the excerpts and their interpretations of participants' and observers' language use that could indicate the relevance of a specific code (MacQueen et al., 1998). If necessary, we also talked through coders' understandings of the definitions of codes applicable to that particular excerpt. If agreement was not reached through this process, coders posed persisting disagreements to the two co-principal investigators of the larger study, as they were able to offer the unique perspective of experienced qualitative researchers who had collected data but had not participated in coding. In the calibration exercises, these team discussions sometimes led to adjustments of the codebook structure and definitions (Campbell et al., 2013), which furthered coders' abilities to apply codes accurately to the data.

Revision of Codebook and Retroactive Coding. Regarding the codebook structure, we occasionally added child and grandchild codes to existing parent codes. Similarly, discussing specific examples in the data allowed us to revise code definitions according to team understanding within a range of applications. Table 8 depicts one such instance in which we worried that one code was defined too broadly and therefore narrowed its scope.

As we made these changes, we simultaneously coded retroactively, making changes to code applications on previously completed documents if necessary. We also found that an important consideration in applying the codebook consistently was establishing rules for how to account for context in the data (as we did when establishing our rules of unitization, discussed above). For example, it was only natural in the context of our study that mentions of lesson elements re-occurred throughout a classroom observation form. Initially, we coded each occurrence. However, this led to a great deal of over-coding that did not contribute new understanding to classroom phenomena. Therefore, we decided to consistently use context to code, but to apply codes only if an excerpt was communicating something significant that had not yet been mentioned (e.g., if a coder knew in one observation that the class was working on a worksheet, that coder could just apply the appropriate code once, rather than to every single excerpt, unless new information was offered).

Bringing It All Together. Following these rules, the team calibrated coding on two additional observation forms to ensure we had sufficiently discussed each code in the codebook. Given the results of this additional calibration, we felt prepared at this point to code observations individually while simultaneously completing these same calibration exercises with a teacher interview, an administrator interview, and a descriptive narrative. These methods of calibration allowed for continued consensus of code application.

Individual Coding

Once these stages of calibration were complete, the coders were each randomly assigned approximately one quarter of the whole dataset, divided roughly equally among data type, which amounted to approximately 80 total documents for each coder. Coders who also collected data were careful to avoid coding observations or interviews they conducted. If needed, coders could reference a coding protocol to troubleshoot common software issues and access written instructions of the established processes described above. As we continued to meet to finalize calibration on interviews and narratives, we also continued to code our assigned observation data individually and resolve issues identified by coders there (MacQueen et al., 1998).

We employed two mechanisms to track and resolve these issues during twice weekly resolution meetings (which were separate from calibration meetings). First, coders had access to a shared document where they could pose global questions cutting across pieces of data. We addressed any new questions to this document during each resolution meeting, though as coders familiarized themselves with the codebook over time, there were fewer global questions to address. Second, as mentioned above, coders had the option to apply the Not Sure code alongside other applicable codes if they had questions pertinent to a specific excerpt. A coder who applied this code was also asked to use Dedoose's built-in memoing tool to attach a brief explanation of the basis for uncertainty to the excerpt. Examples of reasons for applying Not Sure included coders needing help deciding what codes to apply, wanting to discuss whether the threshold for code application was met in an excerpt, pointing out an anomaly in the data or codebook, or having questions on unit length. During the resolution meetings, the team went through the accumulated Not Sure excerpts, agreed on a resolution to each, and then a team leader adjusted code applications to the excerpt in Dedoose as necessary.

Resolving the Not Sure code applications as a team ensured all excerpts were coded and none remained unclear. This process also had numerous other benefits. Primarily, it ensured coders were not drifting from the codebook. It also alerted individual coders to issues that came up for others and enabled them to flag and resolve those issues before anyone else could be thwarted by them. For example, early on in the individual coding process, coders discovered other, less common types of non-examples not previously anticipated. Discussing how to address a new non-example reinforced the original logic behind the codebook and allowed the team to set up new rules to preclude further confusion. Another benefit was the continued development of coders' understandings of complex codes and the opportunity for the team to provide support for one another in the coding process.

When approximately half of the dataset had been coded, we did another co-coding, calibration-type exercise to ensure that coding consistency was maintained over time. We followed the same procedures from the calibration process, having all coders blindly code a single observation form. Again using a spreadsheet to compare individual code applications, we found improved agreement across coders (see Table 5). We also conducted an additional exercise to check for intra-coder consistency (i.e., consistency of one coder's code applications across time). We hid the spreadsheet columns containing the codes applied by each coder on an excerpt more than a month prior, and in real-time, the coders read the excerpt and wrote down the codes they would apply. We then revealed the original code applications and coders checked that they applied the same codes as they had when coding the data previously. Again, we found high within-coder agreement. Satisfied with our across- and within-coder consistency, the team proceeded with coding the rest of the dataset individually, resolving Not Sure excerpts as needed until coding was complete.

Benefits of Our Processes

We see the coding methods described above contributing to related literature and qualitative research on the whole from several vantage points. First, and most basically, our descriptions answer the call for researchers to be more transparent in relaying the steps taken in their treatment of qualitative data (e.g., Burla et al., 2008; Church et al., 2019; Hill et al., 2005; Hruschka et al., 2004; O'Connor and Joffe, 2020; Roberts et al., 2019). A corollary benefit of this transparency is our use of a team-based, consensus process (Hill et al., 1997, 2005). Our overall goal in using this approach was to achieve a balance between utilizing coders' subjectivities and not allowing those subjectivities to color what was actually represented in the data (Hill et al., 2005). More specifically, our use of multiple coders enabled our individual subjectivities to work in our favor when attempting to "unravel the complexities and ambiguities of the data" (Hill et al., 2005: 197) and reduced the potential dangers related to bias present when only one person codes (Hill et al., 2005). Church et al. (2019) have contended that working with multiple coders is worth the challenge, as it will increase the quality of the research due to in-depth discussions among coders, and we agree. What we arrived in the end was not only a way to communicate consistency in our coding that did not hinge on a certain quantifiable level, but also a justifiable way to engage in single-coding the remaining data in a time- and cost-effective manner (Burla et al., 2008; Saldaña, 2016). Thus, our methods are viable for researchers who want to ensure the rigor of their work without double-coding all data or calculating a coefficient (Fereday and Muir-Cochrane, 2006).

Additional advantages of our methods are the innovative ways in which we simultaneously used inductive and deductive approaches to our thematic analysis and applied them via a multilayered codebook to different types of data (cf. Hruschka et al., 2004). We came across no other previous work in which a coding process was depicted that established coding consistency via a team-based approach while mixing inductive and deductive coding schemes *and* dealing with different forms of data in a single dataset. Therefore, our methods offer an innovative way to code across data sources in a manner that is trustworthy and considerate of both the themes that emerge from the data and the existing frameworks initially put in place for analysis.

A final benefit of our approach is that it eschews the more traditional measure of IRR, which arises from a test in which coders are presented with segments of data and asked to apply relevant codes to them that then get measured against a "master" set of codes for each excerpt

(Gwet, 2014).[2] As we discovered when we attempted to take a more traditional reliability test at the onset of our coding, it is not well-suited for a codebook like ours with multiple layered codes that rely on one another for any meaningful use (cf. Burla et al., 2008). Thus, our method intentionally accounts for the context in which the data are situated by using it as a resource for, not hindrance to, interpretation of what it means, and it also enables discussion among coders of how layered codes might interact to provide meaningful interpretations of the phenomena they are trying to understand more deeply (Church et al., 2019).

Conclusions

Despite an increase in qualitative studies, researchers have yet to agree on a standard for determining or evaluating the reliability or consistency of qualitative coding (Armstrong et al., 1997; Hayes and Krippendorff, 2007; MacPhail et al., 2016). Gwet (2014) has posited that this discord exists because a coefficient representing this construct is meaningless without a clear understanding of the researchers' procedures taken to arrive at it (see also Hammer and Berland, 2014; Schoenfeld, 1992). Saldaña (2016: 37) has recommended such procedures include "intensive group discussion . . . coder adjudication, and simple group consensus," as ours did. Thus, we have used our study of gifted education programs as a vehicle through which to explicate the interactive processes we followed to achieve consistency in coding our data.

Though we use our study to illustrate our practices, we believe our methods and the lessons we learned from them advance the conversation surrounding coding in qualitative inquiry both within and outside of the field of education. The means of data collection we used in our study are not specific to educational research, but are are commonly employed in many if not all

[2] For more comprehensive discussion of the downfalls of traditional IRR tests, which bears on our methods but is outside of the scope of our manuscript, we refer readers to Kurasaki (2000) and Morse (1999).

types of qualitative research (at least to some degree). As a result, our methods of coding said data can be utilized broadly across social science research. While studies in different fields will of course have unique analytic goals, coding of qualitative data must be done in a systematic and transparent way if the processes (and the conclusions reached) are to be trusted. We offer one such trustworthy way for completing the coding process and preparing researchers to draw sound conclusions from their coding that are applicable and relevant to their fields.

What is more, scholars have offered even less support for establishing consistency in coding while reconciling deductive and inductive schemes or dealing with different forms of data in a single dataset. By elucidating how we contemporaneously built a layered codebook, established unitization of the data, trained coders, and coded our data, we have taken a step to remedy this discord by offering new insights into a defensible process of ensuring consistency and trustworthiness in qualitative research. Our methodology brings these disparate constraints and imperatives in qualitative work together while simultaneously prioritizing the contextual and interpretive characteristics of this type of inquiry in a way that contributes to improved future qualitative research.

Recommendations

The methods described here offer an innovative approach to coding qualitative data, but that is not to say that we carried them out without complication. Thus, we end by addressing these challenges to facilitate future use of these methods, as we believe that just as our methods can be applied across the social sciences, so too can the lessons we learned throughout the coding process.

We have discussed how we were able to recreate our methodology here due to steadfast documentation of our processes as they occurred. This was through the use of coding questions protocols, meeting agendas, and a methodological journal. We have also thoroughly shared how we iteratively created, refined, and tested our codebook. While we kept a detailed record of codebook changes from the beginning to the end of the process, in hindsight we would have also liked to document our thinking more thoroughly as we edited codes to contextualize our decisions should we need to return to the codebook down the road. Though this realization pertains more to our needs in our study than to the larger description of our methods, we share it here so that others might consider the need for it in their own application of our methods.

Regarding the coding itself, we have discussed the benefits of utilizing a team-based approach. However, the literature has also stressed the importance of adhering to certain criteria for quality within this approach, particularly when a team is large in number and made up of individuals with varied backgrounds (e.g., Charmaz, 2006; Church et al., 2019; Hill et al., 1997, 2005; Smith and McGannon, 2018). Our team met these criteria, as it was comprised of individuals with different levels of methodological expertise and diverse areas of specialization, both within and outside the field of education. While these differences provided us with a breadth of perspectives, we also had to be acutely aware of the undesirable power dynamics they could create in our discussions and the potential danger of diverse perspectives turning into biases that could permeate our coding (Charmaz, 2006; Church et al., 2019; Smith and McGannon, 2018). We addressed these issues by openly sharing our interpretations of codes and excerpts, which were divergent at times, and by considering how they might be informed by our own experiences. We also worked consciously in our discussions to combat groupthink by encouraging team members to share their logic freely and by being open to code applications that were perhaps only initially proposed by one of the coders; similarly, we aimed to avoid the designation of any one person's interpretations or code applications as "expert" by using our diverse perspectives to come to group consensus on what a code signified (Hill et al., 2005: 198; Hill et al., 1997). We recommend that all qualitative teams that consider using the methods described here remain aware of these elements, which can turn into constraints if not properly addressed throughout the coding procedure.

Finally, we would be remiss not to emphasize that the way we present our methods on the page cannot fully capture neither the depth and breadth, nor the iterative and reflexive nature, of our processes. On the whole, our processes took approximately five months from the beginning of coding to the end, with the team holding 47 meetings in this time period (approximately an average of two meetings per week). The extended time for implementation of this process could be for several reasons. We believe it is partially due to our meticulous approach to coding, and to our decision at the onset of the coding process to code not only with our research design in mind, but also in anticipation of what might emerge as noteworthy for future inquiry, which undoubtedly added time to our processes. However, we posit that the main factor in how long coding took can be attributed to elements particular to our study, such as: amount of data, number of coders, coders' levels of comfort with the codebook, frequency/length of coding meetings, and number of times coders believed that a real-time coding exercise would be beneficial. Such elements would be unique to every study in which our methods are used. We were fortunate to have four coders who were intimately acquainted with both the data and codebook to handle our large dataset, which undoubtedly contributed to the shortening of meeting lengths over time. We believe that the level of detail we achieved in our coding is worth striving for in any study and thus see immense value in our methods because of the ways they integrate trademark elements of qualitative research. Readers who wish to use our methods should understand that the overall amount of time it may take to implement these processes depends more than anything on the individual characteristics of their particular study.

Similarly, there was much necessary overlap across codebook development, calibration, individual coding, and the discussions that surrounded these components of the process, and they by no means followed a linear trajectory (Fereday and Muir-Cochrane, 2006). Such is the nature of most branches of qualitative inquiry. We must also point out our processes did not lead to a fossilized codebook or a bias-free transmission of the code definitions into the minds of the coders. While some codes proved easier to define clearly and interpret with more certainty than others, as we have moved from the coding stage toward analysis, elements of uncertainty still arise. We address these issues by continuing the processes of open discussion outlined above and reminding ourselves that such challenges are inevitable – but also manageable – in qualitative research. Therefore, we present our methods here not as the ultimate, error-proof answer to the issues that may arise during qualitative coding, but rather as a way of using hallmark characteristics of qualitative work to ensure consistency when team-coding across varied types of data deductively and inductively.

References

Aguinis H, Hill N and Bailey J (2019) Best practices in data collection and preparation: Recommendations for reviewers, editors, and authors. *Organizational Research Methods* 1–16.

Armstrong D, Gosling A, Weinman J and Marteau T (1997) The place of inter-rater reliability in qualitative research: an empirical study. *Sociology* 31(3): 597–606.

Bernard HR (2000) *Social research methods: Qualitative and quantitative approaches.* Thousand Oaks, CA: SAGE.

Braun V and Clarke V (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2): 77–101.

Brinkmann S, Jacobsen, MH and Kristiansen, S (2014) Historical overview of qualitative research in the social sciences. In: Leavy P (ed), *The Oxford Handbook of Qualitative Research.* Oxford, UK: Oxford University Press, pp. 17–42.

Burla L, Knierim B, Barth J, Liewald K, Duetz M and Abel T (2008) From text to codings: Intercoder reliability assessment in qualitative content analysis. *Nursing Research* 57(2): 113–117.

Campbell JL, Quincy C, Osserman J and Pedersen OK (2013) Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research* 42(3): 294–320.

Carey JW, Morgan M and Oxtoby MJ (1996) Intercoder agreement in analysis of responses to open-ended interview questions: Examples from tuberculosis research. *CAM Journal* 8(3): 1–5.

Charmaz K (2006) *Constructing grounded theory: A practical guide through qualitative* *analysis.* Thousand Oaks, CA: SAGE.

Church SP, Dunn M and Prokopy, L (2019) Benefits to qualitative data quality with multiple coders: Two case studies in multi-coder data analysis. *Journal of Rural Social Science* 34(1/2): 1–14.

Eisner E (1998) *The enlightened eye: Qualitative inquiry and the enhancement of educational practice.* Upper Saddle River, NJ: Prentice Hall.

Elliott R, Fischer CT and Rennie DL (1999) Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology* 38(3): 215–229.

Elmore PB and Woehlke PL (1998) Twenty years of research methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research. In: *American Educational Research Association Annual Meeting*, San Diego, CA.

Fereday J and Muir-Cochrane E (2006) Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods* 5(1): 80–92.

Garrison DR, Cleveland-Innes M, Koole M and Kappelman J (2006) Revisiting methodological issues in transcript analysis: Negotiated coding and reliability. *The Internet and Higher Education* 9(1): 1–8.

Gergen KJ, Josselson R and Freeman M (2015) The promises of qualitative inquiry. *American Psychologist* 70(1): 1-9.

Glazer JL and Egan C (2018) The ties that bind: Building civic capacity for the Tennessee Achievement School District. *American Educational Research Journal* 55(5): 928–964.

Golafshani N (2003) Understanding reliability and validity in qualitative research. *The Qualitative Report* 8(4): 597–606.

Guba E and Lincoln Y (1994) Competing paradigms in qualitative research. In: Denzin NK and Lincoln Y (eds), *Handbook of Qualitative Research*. Thousand Oaks, CA: SAGE, pp. 105–17.

Gwet KL (2014) *Handbook of inter-rater reliability* (4th ed.). Gaithersburg, MD: Advanced Analytics, LLC.

Hammer D and Berland LK (2014) Confusing claims for data: A critique of common practices for presenting qualitative research on learning. *Journal of the Learning Sciences* 23(1): 37–46.

Hayes AF and Krippendorff K (2007) Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1(1): 77–89.

Hill C, Knox S, Thompson B, Williams E, Hess S and Ladany N (2005) Consensual qualitative research: An update. *Journal of Counseling Psychology* 52(2): 196–205.

Hill C, Thompson B and Williams E (1997) A guide to conducting consensual qualitative research. *The Counseling Psychologist:* 25(4): 517–572.

Hodson R (1998) Organizational ethnographies: An underutilized resource in the sociology of work. *Social Forces* 76(4): 1173–1208.

Hruschka DJ, Schwartz D, St. John DC, Picone-Decaro E, Jenkins RA and Carey JW (2004) Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field Methods* 16(3): 307–331.

King N (2004) Using templates in the thematic analysis of text. In: Cassell C and Symon G (eds), *Essential Guide to Qualitative Methods in Organizational Research.* London, UK:

SAGE, pp. 257–270.

Kirk J and Miller M (1986) *Reliability and validity in qualitative research.* Newbury Park, CA: SAGE.

Kurasaki K (2000) Intercoder reliability for validating conclusions drawn from open-ended interview data. *Field Methods* 12(3): 179–194.

Levitt H, Bamberg M, Creswell J, Frost D, Josselson R and Suárez-Orozco C (2018) Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA publications and communications board task force report. *American Psychologist* 73(1): 26–46.

Lincoln Y and Guba E (1985) *Naturalistic inquiry*. London, UK: SAGE.

Lombard M, Snyder-Duch J and Bracken CC (2002) Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research* 28(4): 587–604.

MacPhail C, Khoza N, Abler L and Ranganathan M (2016) Process guidelines for establishing intercoder reliability in qualitative studies. *Qualitative Research* 16(2): 198–212.

MacQueen KM, McLellan E, Kay K and Milstein B (1998) Codebook development for team-based qualitative analysis. *CAM Journal* 10(2): 31–36.

Marshall C and Rossman G (1989) *Designing qualitative research*. Newbury Park, CA: SAGE.

Mays N and Pope C (2000) Assessing quality in qualitative research. *The BMJ* 320(7226): 50–52.

Miles M and Huberman A (1994) *Qualitative data analysis*. Thousand Oaks, CA: SAGE.

Morse J (1999) Myth #93: Reliability and validity are not relevant to qualitative inquiry. *Qualitative Health Research* 9(6): 717–718.

Morse J, Barrett M, Mayan M, Olson K and Spiers J (2002) Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods* 1(2): 13–22.

Noble H and Smith J (2015) Issues of validity and reliability in qualitative research. *Evidence Based Nursing* 18(2): 34–35.

Nowell L, Norris J, White D and Moules N (2017) Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods* 16(1): 1-13.

O'Connor CO and Joffe H (2020) Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Research* 19:1–13.

Peterson J (2019) Presenting a qualitative study: A reviewer's perspective. *Gifted Child Quarterly* 63(3): 147–158.

Roberts K, Dowell A and Nie, J-B (2019) Attempting rigour and replicability in thematic analysis of qualitative research data: A case study of codebook development. *BMC Medical Research Methodology* 19(66): 1–8.

Saldaña J (2016) *The coding manual for qualitative researchers* (3rd ed.). London, UK: SAGE.

Sandelowski M (1986) The problem of rigor in qualitative research. *Advances in Nursing Science* 8(3): 27–37.

Schoenfeld AH (1992) On paradigms and methods: What do you do when the ones you know don't do what you want them to? Issues in the analysis of data in the form of videotapes. *The Journal of the Learning Sciences* 2(2): 179–214.

Silverman D (2000) *Doing qualitative research: A practical handbook*. London, UK: SAGE.

Smith ML (1987) Publishing qualitative research. *American Educational Research Journal* 24: 173–183.

Smith B and McGannon K (2018) Developing rigor in qualitative research: Problems and opportunities within sport and exercise psychology. *International Review of Sport and Exercise Psychology* 11(1): 101–121.

Stearns E, Bottia MC, Giersch J, Mickelson RA, Moller S, Jha N and Dancy M (2019). Do relative advantages in STEM grades explain the gender gap in selection of a STEM major in college? A multimethod answer. *American Educational Research Journal*.

Syed M and Nelson S (2015) Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood* 3(6): 375–387.

Tobin G and Begley C (2004) Methodological rigour within a qualitative framework. *Journal of Advanced Nursing* 48(4): 388–396.

Tracy S (2010) Qualitative quality: Eight “big-tent” criteria for excellent qualitative research. *Qualitative Inquiry* 16(10): 837–851.

Weston C, Gandell T, Beauchamp J, McAlpine L, Wiseman C and Beauchamp C (2001) Analyzing interview data: The development and evolution of a coding system. *Qualitative Sociology* 24(3): 381–400.

Table 1

Example of Code Application Based on Interview Prompt and Additional Relevant Information

Excerpt	Coder 1	Coder 2	Coder 3	Coder 4	Memos	Finalized Codes
I: Okay. How do you evaluate your students or how do they demonstrate their learning? P: Verbal participation, classwork, classroom assessment and biweekly assessment which we document and track the data using a performance matters district website.	Assessment	**Instruction** **Thinking Levels** Assessment *Pre/Formative Assessment*	Assessment *Pre/Formative Assessment* Summative Assessment	Assessment *Pre/Formative Assessment* Summative Assessment **Classroom Climate** **Teacher Expectations**	NA	Assessment Pre/Formative Assessment Summative Assessment Classroom Climate Teacher Expectations

Note. This table displays one excerpt in a spreadsheet of code applications to a piece of data. Each column represents a coder, each row comprises a unit of data (excerpt), and each cell displays the codes that each of the four coders applied to this excerpt. The code applications were then coded to show 100% (plain text), 75% (italic text), 50% (underlined text), and 25% (bold text) agreement across coders on the coding of this excerpt. This same text coding scheme applies to Tables 4-7.

Table 2

Types of Non-examples in Structured Observation Data

Non-Example Type	**Response to Prompt: Creates opportunities for students to examine big ideas, essential questions, concepts, and/or principles.**
1) Empty	Quality: Frequency:
2) Note of "no evidence"	No evidence. Quality: 0 Frequency: 0
3) Rejection of prompt	Big idea = fractions? No evidence of an essential question anywhere (verbal or written on board). Quality: 0 Frequency: 0
4) Negative example	No, as described above and in open-ended comments she even discourages the attempt on a student's part to find an alternative way to solve a problem once a pattern has been discerned. Teacher does ask for synonyms for patterns, but she doesn't even ask for examples from their lives or experiences in math (e.g., 2, 4, 6, 8 or 1, 3, 5, 7). Or to look for examples of patterns in their world. Quality: 0 Frequency: 0
5) Reference to interview (by observer)	Not observed in this lesson. Noted in interviews that she uses essential questions but not seen today. Quality: 0 Frequency: 0
6) Description unrelated to prompt	Students will define each format on their folded paper; open up and there will be a problem; solve problem. Quality: 1 Frequency: 1

Note. "Quality" and "Frequency" were categories on our observation protocol used as measurements of elements that our theoretical framework primed us to look for and expect to see in the classroom. They were not codes in our codebook, but rather elements that we took into consideration when coding for non-examples.

Table 3A

Example of Code Application to Semi-Structured Excerpt with No Comments

Excerpt	Code(s)
Advanced Content (AC)	A. Advanced Content A. Advanced Content – no

Note. The observer did not note any advanced content and left the response area blank, leading to the applied codes.

Table 3B

Example of Code Application to Semi-Structured Excerpt with Comments

Excerpt	Code(s)
Advanced Content (AC) No. Students were working on writing standard essays with introduction, three main points, and conclusion from the grade level curriculum.	A. Advanced Content A. Advanced Content – no Curriculum (The What) Instruction (The How) Thinking Levels Classroom Climate Student Engagement

Note. The observer did not note any advanced content, but described what took place instead. Therefore, although this excerpt was coded as A. Advanced Content – no, it also received codes for the information the observer did provide.

Table 3C

Example of Code Application to Structured Excerpt with Ratings of Zero

Excerpt	Code(s)
1. **Gives appropriate additional support to students as needed. (re-reading or rephrases instructions and materials to make sure students understand; adding extra material to lesson, appropriate use of technology, answering questions- if there is no evidence that this kind of support is necessary, item should be marked, N/A) Comments:** Student who was absent for the prior lesson is not given any information to be brought up to speed. Perhaps the teacher knows this student is capable of catching on without that help? Teacher does start to call on her and notes the absence and moves to another student but that doesn't provide any support. **Q:** 0 **F:** 0	1. Gives appropriate additional support 1. Gives appropriate additional support – no Teacher Expectation Classroom Climate

Note. The observer did not observe the teacher use additional supports and thus rated the quality (Q) and frequency (F) of such supports both as 0. However, the observer did describe a missed opportunity for providing additional support in the comments, prompting coders to code for Teacher Expectation and Classroom Climate.

Table 3D

Example of Code Application to Structured Excerpt with Ratings and Comments

Excerpt	Code(s)
1. **Gives appropriate additional support to students as needed. (re-reading or rephrases instructions and materials to make sure students understand; adding extra material to lesson, appropriate use of technology, answering questions- if there is no evidence that this kind of support is necessary, item should be marked, N/A) Comments:** Yes, teacher answers questions as needed. **Q:** 1 **F:** 1	1. Gives appropriate additional support Instruction (The How) Instructional Support

Note. The observer noted the teacher's use of additional support to students both in the comments and in rating the quality (Q) and frequency (F) both as 1. The excerpt was then coded not only for the affirmative 1. Gives appropriate additional support based on these ratings, but also for Instruction and Instructional Support based on the comments.

Table 4

Example of Code Application with Mixed Levels of Coder Agreement (Early in Coding Process)

Excerpt	Coder 1	Coder 2	Coder 3	Coder 4	Memos	Finalized Codes
I: Okay. So then when you are the one making the decisions and choosing instructional strategies what kinds of factors guide those decisions for you? P: Student needs, student knowledge of learners, knowledge of the population you're working with, I work with an ELL population about twenty-five to thirty-five percentile is my first group, my a.m. group, so their needs are going to be different than my afternoon with my ESL, they're working at grade level for the most part. My core curriculum will be the same, but the presentation	*Curriculum* Instruction Adjustments to Instruction **Student Challenge** **Instructional Support** *Differentiation* *Differentiation across Classrooms* Assessment Pre-/Formative Assessment **Assessment with Adjustment** Classroom Climate Teacher Expectations School Context/Climate	Instruction Adjustments to Instruction **Thinking Levels**	*Curriculum* Pacing Guide **Deviating** Instruction Adjustments to Instruction *Differentiation* *Differentiation across Classrooms* School Context/Climate	*Curriculum* Pacing Guide **Following** Instruction Adjustments to Instruction *Differentiation* *Differentiation across Classrooms* Assessment Pre-/Formative Assessment Classroom Climate Teacher Expectations	N/A	Curriculum Instruction Adjustments to Instruction Differentiation Differentiation across Classrooms Assessment Pre-/Formative Assessment Assessment with Adjustment Classroom Climate Teacher Expectation School Context/Climate

will be slightly different, again talking about the multi-sensory and the multi-modes of learning, and the pacing, so I do move a little slower with the morning class. I have to in order for me to be able to evaluate that they're mastering and understanding the concept.						

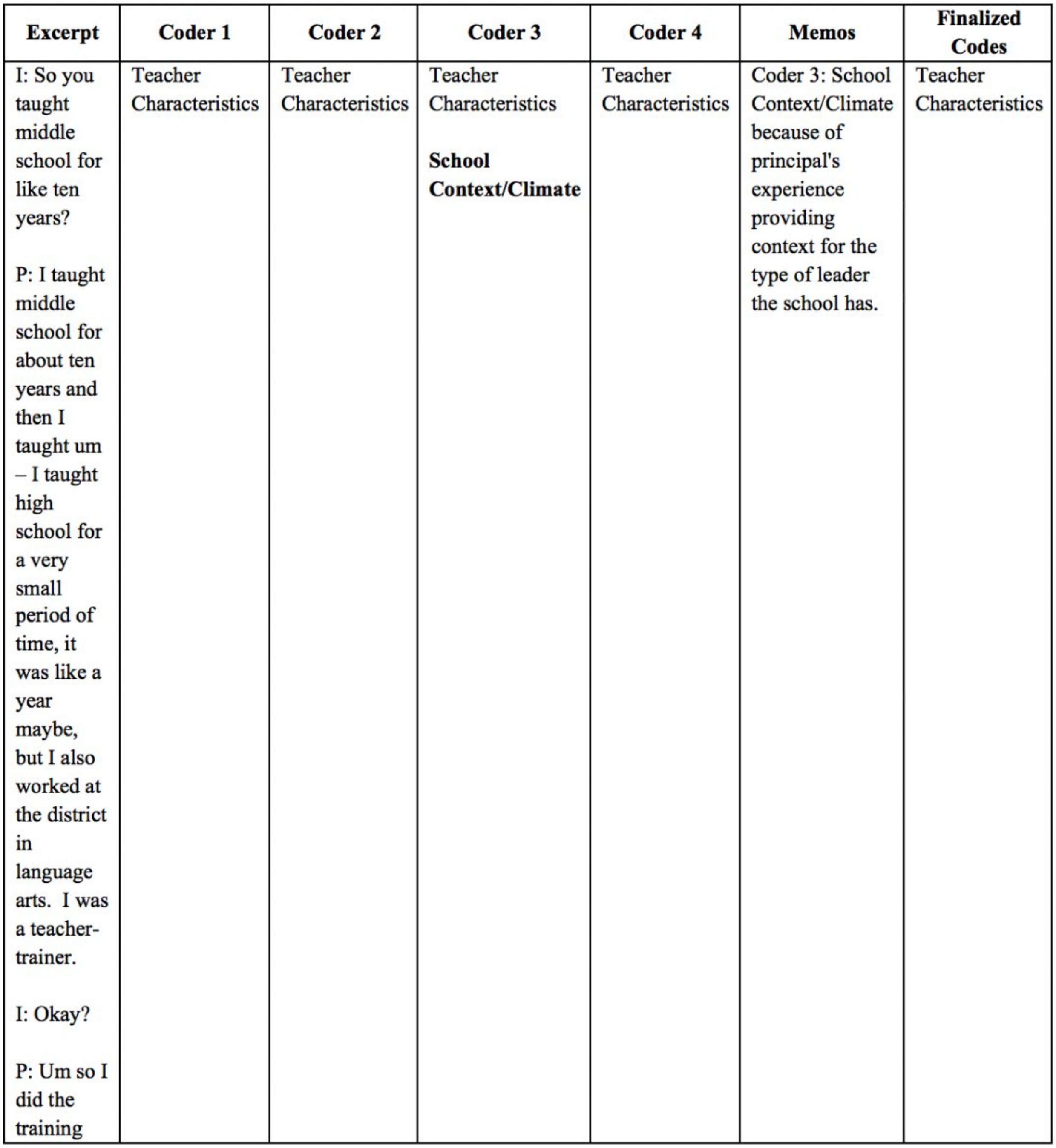
Table 5

Example of Code Application with Higher Levels of Coder Agreement (Later in Coding Process)

Excerpt	Coder 1	Coder 2	Coder 3	Coder 4	Memos	Finalized Codes
The lesson flowed easily from teacher-guided discussion, small group discussions, to teacher-student and student-student discussions. Talk moves facilitated the interactions. It was evident that the teacher and students were accustomed to working in small groups and using talk moves. Later, another teacher noted that small group discussions were an expectation set by the administration.	Instruction Grouping Classroom Climate Student Engagement Student Interactions Teacher Expectations *School Context/Climate* *Educator Expectations Other*	Instruction Grouping Classroom Climate Student Engagement Student Interactions Teacher Expectations *Educator Expectations Other*	Instruction Grouping Classroom Climate Student Engagement Student Interactions Teacher Expectations *School Context/Climate* *Educator Expectations Other*	Instruction Grouping Classroom Climate Teacher Expectations Student Interactions Student Engagement *School Context/Climate* (Not Sure)	Coder 4: Did anyone code talk moves as Teacher Expectations?	Instruction Grouping Classroom Climate Student Engagement Student Interactions Teacher Expectations School Context/Climate Educator Expectations Other

Table 6

Example of 25%: When One Coder's Applied Code Was Not Adopted by the Team



Excerpt	Coder 1	Coder 2	Coder 3	Coder 4	Memos	Finalized Codes
I: So you taught middle school for like ten years? P: I taught middle school for about ten years and then I taught um – I taught high school for a very small period of time, it was like a year maybe, but I also worked at the district in language arts. I was a teacher-trainer. I: Okay? P: Um so I did the training	Teacher Characteristics	Teacher Characteristics	Teacher Characteristics **School Context/Climate**	Teacher Characteristics	Coder 3: School Context/Climate because of principal's experience providing context for the type of leader the school has.	Teacher Characteristics

for the teachers and so I worked there and then I was also a reading coach and a reading leader, and then I worked at the regional office as a curriculum support for students, so.						

Table 7

Example of 25%: When One Coder's Applied Code Was Adopted by the Team

Excerpt	Coder 1	Coder 2	Coder 3	Coder 4	Memos	Finalized Codes
I: So do you have any um – are you privy to any information about how this particular program was chosen? P: No, um the only thing I do know is based on the number of student, we're funded based on the number of kids that we have and so if you have a hundred kids you get a certain number of teachers based on the number of students. That's why also it's important to me to push for us to get more kids tested and really	Program Decision-Making *Identification*	*School-Level Programming*	*School-Level Programming* *Identification* **School Context/Climate**	Program Decision-Making *Identification* *School-Level Programming*	N/A	School-Level Programming Identification Program Decision-Making School Context/Climate

identify; I'm sure they're there; we just have to identify them more accurately.						

Table 8

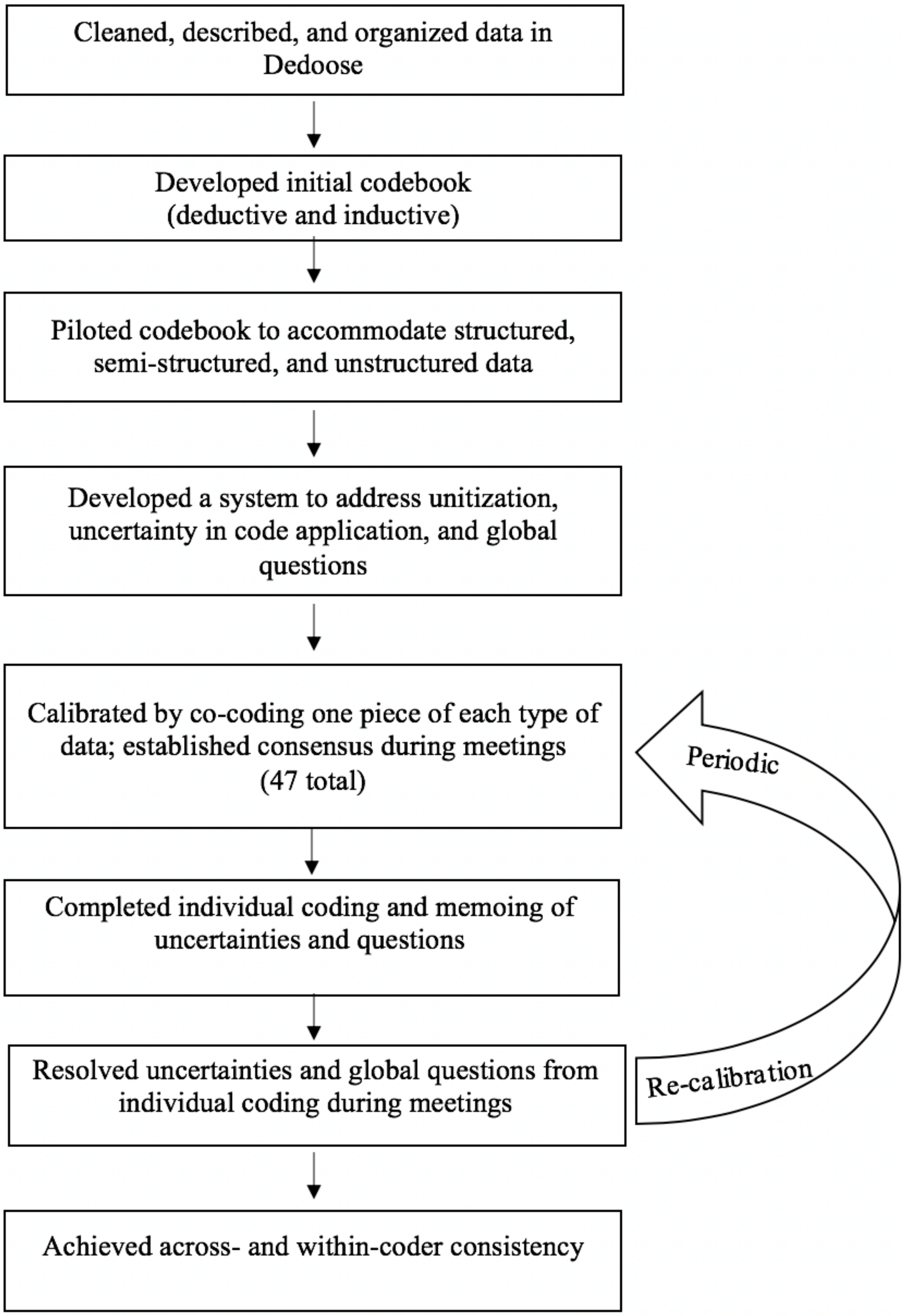
Example of Code Revision and Codebook Evolution

Code	Definition
Thinking Levels	Questions and/or tasks that engage students in various levels of thinking from lower levels (applying, calculating, defining, remembering, restating, solving, understanding) to higher levels (analyzing, creating, problem-solving, comparing/contrasting, evaluating, defending, metacognition, discovery/inquiry). Use whether or not you see a student actually doing/thinking something (i.e., if the teacher asks the students to do/think, but the observer doesn't actually note whether or not the students did/thought it). **Describes what students are asked to actively do with the content, not how hard/easy the teacher perceives the content to be. Look for "action word" of what students are being asked to do with content.**

Note. The plain text represents the original definition, and the bold text represents additions to the definition as the codebook was communally revised.

Figure 1

Overview of Methods of Coding Consistency


Cleaned, described, and organized data in Dedoose
Developed initial codebook (deductive and inductive)
Piloted codebook to accommodate structured, semi-structured, and unstructured data
Developed a system to address unitization, uncertainty in code application, and global questions
Calibrated by co-coding one piece of each type of data; established consensus during meetings (47 total)
Periodic
Completed individual coding and memoing of uncertainties and questions
Resolved uncertainties and global questions from individual coding during meetings
Re-calibration
Achieved across- and within-coder consistency