2020

Monitoring implementation in program evaluation with direct audio coding

Jennifer Farley
University of Nebraska-Lincoln, jfarley3@unl.edu

Kristin Duppong-Hurley
University of Nebraska–Lincoln, kristin.hurley@unl.edu

A. Angelique Aitken
University of Nebraska-Lincoln, aaitken2@unl.edu

Monitoring implementation in program evaluation with direct audio coding

Jennifer Farley, Kristin Duppong Hurley, & A. Angelique Aitken

University of Nebraska-Lincoln

Corresponding author — J. Farley, 212 Barkley Memorial Union, University of Nebraska-Lincoln, Lincoln, NE, 68583
Email — J. Farley, jfarley3@unl.edu; K. Duppong Hurley, Kristin.hurley@unl.edu; A. A. Aitken, aaitken2@unl.edu

Abstract
This project explored the reliability and utility of transcription in coding qualitative data across two studies in a program evaluation context. The first study tested the method of direct audio coding, or coding audio files without transcripts, using qualitative data software. The presence and frequency of codes applied in direct audio coding and traditional transcription coding were compared and the two methods produced similar results. Direct audio coding was then employed in an evaluation study to monitor implementation and the method and to be reliable. Implications are discussed with considerations for both researchers and practitioners.
Keywords: Transcription, Implementation, Direct audio coding, Program evaluation

1. Introduction

In program evaluation, qualitative data can offer valuable information about the perspectives and experiences of research participants (Neal, Neal, VanDyke, & Kornbluh, 2015). While such information clearly

Published in *Evaluation and Program Planning* 83 (2020) 101854
DOI: 10.1016/j.evalprogplan.2020.101854

Submitted 8 November 2019; revised 4 March 2020; accepted 30 July 2020.

benefits the evaluation, the processes by which qualitative data are collected, managed, and analyzed are less clear and may vary according to the research design and questions. Transcription, or the generation of type-written text from an audio file (Halcomb & Davidson, 2006; Tracy, 2013), is frequently used to manage qualitative data because it creates a complete and detailed verbal record, which allows for a close review of the data by working with the actual text from the conversations (Tracy, 2013). While such transcripts can be generated and coded using qualitative data analysis software, advances in technology allow for coding audio and video files directly, thus making it possible to skip the transcription process entirely. While coding directly from audio and video files still allows for the ability to review the original words of the respondent, as is possible from transcripts, it eliminates the extra step of producing the transcript. However, limited research exists which compares coding of audio files and transcripts, especially when used in implementation studies examining the presence or absence of content within service-delivery sessions. Therefore, it is unknown if coding audio files directly would produce the same results as coding transcripts of sessions when identifying topics included in service delivery sessions. It is also uncertain if different elements may stand out more when written in transcribed text than heard in an audio recording of such a session.

While the literature calls for increased use of qualitative methods in program evaluation (Christie & Fleischer, 2010), there are several drawbacks to using routine qualitative methods such as transcription. Transcription is a time-consuming process, (Neal et al., 2015; Skillman et al., 2018; Tessier, 2012), which can be made longer if the recording is of low quality or if the individuals speaking are difficult to understand (Tracy, 2013). This lengthy process can also be expensive (Neal et al., 2015; Skillman et al., 2018; Tessier, 2012), as services of a professional transcriptionist can cost $100 an hour or more (Tracy, 2013). The costs associated with transcription services typically make up a large portion of a study's budget and may determine the number or length of interviews conducted (Crichton & Kinash, 2003). Given that verbal and written communication use different structures and syntaxes, written transcripts may omit data or include altered sentence structures, mistaken words, and improper use quotation marks (Poland, 1995). Furthermore, transcripts may fail to adequately capture participant voice, or other relevant data (Crichton & Childs, 2005; Greenwood, Kendrick,

Davies, & Gill, 2017) present in the audio file. In this way, a transcript "flattens the potentially rich, three-dimensional quality of the original footage into a two-dimensional text format," (Crichton & Childs, 2005, p. 3). However, despite these challenges, there is limited guidance in the literature to support evaluators in making decisions regarding whether or not to use transcription in a given project.

While transcription transforms conversations into usable data, researchers have explored alternate ways to streamline qualitative data collection and analysis because of the disadvantages associated with transcription. Some have suggested that it may not always be necessary to transcribe audio data (Saldaña, 2016; Tracy, 2013) depending on the how the data will be analyzed. Furthermore, use of audio and video benefits the research process by allowing the researcher to hear the participant's voice (e.g., intonation, inflections, pauses, passion) rather than read their words (Crichton & Childs, 2005; Tessier, 2012).

Direct audio coding is the method by which data are coded while listening to an audio file without (or before) transcription. Greenwood et al. (2017) found consistent themes and results when they compared data from transcripts and audio recordings. Other researchers have demonstrated the benefits of using direct audio coding in program evaluations to document functions, monitor processes, and incorporate participant voices (Neal et al., 2015; Tessier, 2012). Some have found that direct audio recording is particularly useful in evaluations where analysis and reporting are time sensitive (Halcomb & Davidson, 2006; Neal et al., 2015).

The increased use of the direct audio coding method over the last few years may be related to an increased use of software programs to analyze qualitative data. The development of Computer Assisted Qualitative Data Analysis Software (CAQDAS) in the mid-1990s, opened new data analysis opportunities to qualitative researchers (Cope, 2014). Using CAQDAS to manage and analyze qualitative data has allowed researchers to conduct more in-depth analyses (e.g., word counts, counting cases, relationships between codes), manage data more efficiently, and collaborate between multiple researchers with ease (Basit, 2003; Cope, 2014; Leech & Onwuegbuzie, 2007; Vander Putten & Nolen, 2010; Woods, Paulus, Atkins, & Macklin, 2016). Woods et al. (2016) conducted a review of how software programs are used in qualitative research. They found that CAQDAS have been used across diverse disciplines to analyze qualitative data collected through a number of methods, including documents, surveys

(open-ended questions), interviews, focus groups, and field notes. However, they found little evidence of researchers employing direct audio coding techniques, with only two of 763 studies indicating the use of software to code "directly from multimedia files" (Woods et al., 2016, p.606). Only one study was located that compared themes identified by direct audio coding and transcription coding, and the results indicated that both methods identified similar themes (Greenwood et al., 2017). Moreover, there are no known applications of the method using CAQDAS in implementation studies of service delivery content.

The purpose of this project was to determine whether direct audio coding was a viable and reliable method to monitor meetings between participants and staff in a program evaluation project. To this end, the direct audio coding method was tested in two ways. First, we conducted a comparison study to examine the level of agreement and reliability reached by raters when using direct audio coding and transcription coding. This first study applied codes, specific to the topics discussed during service delivery, to a sample of audio files (n = 15) using both transcription and direct audio coding methods. We then expanded our inquiry of direct audio coding by examining reliability of the method in monitoring service delivery implementation in a large program evaluation study of an in-home family intervention. In this evaluation study, we used direct audio coding to apply codes, specific to the core components of the program, to a larger sample of audio files (n = 102) for which inter-rater reliability was measured.

2. General methods

2.1. Setting

Both studies were components of a larger, multi-year randomized evaluation of an intensive in-home family intervention program for families of children with emotional and behavioral challenges. The evaluation was approved by the University's Institutional Review Board and the agency offering the intervention. Participating families resided in a Midwestern state and were invited to participate in the study after they had called a family helpline because of their child's behavior. Of the 377 families who provided informed consent, 76 did not complete the required

intake materials, and one did not meet the inclusion criteria. The remaining 300 families were randomly assigned to either the intervention (*n*=152) or the control (*n*=148) group. Families in the intervention group met in-person, for eight to 12 weeks, with a trained and supervised Family Consultant who provided additional education and supports tailored to the family's specific needs regarding their child's behavior. For example, Family Consultant services would help parents to improve parenting skills, understand family functioning, improve family engagement, and access community resources (Duppong Hurley et al., 2019).

2.2. Data collection

Family Consultants recorded up to three sessions with each family (i.e., beginning, middle, end of the intervention) to monitor program fidelity. Password-protected iPads were used by Family Consultants to record program sessions. While the video function was used, to increase comfort of the families the camera was directed toward a wall or laid flat on the table so only audio was collected. After the audio was recorded, the agency downloaded the file, stored it on a secure server, and then deleted the file from the iPad. The agency then provided the recordings to the evaluation team through a shared secure server.

2.3. Data analysis

Implementation was monitored through a thematic analysis of meetings between Family Consultants and program participants. Procedures were established for transcribing recorded sessions, coding transcripts, and direct audio coding. The codebook was established for the larger evaluation study, which included sets of codes based on the intervention's (a) core program components (e.g., relationship building, risk screening, teaching skills, supports and resources), (b) activities (e.g., scripting, modeling, practice and feedback), (c) topics discussed (e.g., physical health care, behavioral/mental health care, substance abuse, child education), and (d) skills developed by participants (e.g., effective praise, consequences, family meetings, routines).

Over the course of the four-year project, the evaluation team's Data Manager trained a team of 24 data assistants (undergraduate and graduate students) in all data analysis procedures, including transcription,

transcript coding, and direct audio coding. The training included: (a) becoming familiar with the codebook and procedures, (b) practice application of codes on an audio and/or transcript, (c) reviewing results with the Data Manager, and (d) repeating steps b and c for seven practice files. Data assistants demonstrated reliability with at least 80 % agreement on three consecutive independent coding assignments for both coding methods before being assigned to either transcribe, code with the transcript, or conduct direct audio coding for a given recorded session. Assignments were made so that the same data assistant did not perform multiple functions on the same recorded session (transcribing, transcript coding, direct audio coding).

3. Comparison study

3.1. Method

For the first study, we selected a random sample of 15 recorded sessions, (16 % of the 241 recordings collected), and implemented both transcription and direct audio coding procedures. Coding by both transcription and direct audio coding is expensive and funds did not exist to dual code the entire sample of recorded sessions. Thus comparing about 15 % of the sessions was reasonable to determine whether or not the direct audio coding process held promise. Data assistants used the qualitative data analysis software NVivo 11 (QSR International, 2016) for all transcribing and coding procedures (i.e., direct audio coding and transcription coding). After coding was complete, we compared results of the two methods. Qualitative software reports and queries detailed the presence, frequency, and agreement for each code, which were compared across coding methods (transcription and direct audio coding). We then calculated differences between the methods and assessed inter-rater reliability with the Kappa coefficient.

3.1.1. Transcription procedures

Data assistants imported 15 recordings into the qualitative software and transcribed them verbatim. The transcripts were created so that each time the individual speaking changed, their dialogue was recorded

on a new numbered line and each line was timestamped. Prior to coding, all transcripts were reviewed and compared to the accompanying recordings. Small edits were made, as needed, to provide a more accurate transcript. During the process of creating the transcripts, the data assistants removed identifying information, such as names of individuals or service providers, and replaced them with standard abbreviations used in all transcripts (e.g., CG for caregiver, Y for youth). Data assistants were trained to transcribe, as well as to code with transcription and direct audio coding methods (see coding procedures). However, data assistants only performed one of these three tasks (transcription, transcription coding, or direct audio coding) for any one recorded session.

3.1.2. Coding procedures

In this initial study, we applied codes specific to topics discussed during the intervention service delivery. The topic codes require analysis of *what* is discussed between the Family Consultant and participant. Specifically, the following four topics were coded; substance abuse, child education, child's behavioral/mental health, and physical health. Data assistants worked from one master copy of the project located on the server, and all codes were established within the project. Codes were applied for the entire length of time the topic was discussed in the recording. While these codes typically apply to large segments of the audio/transcript, there were portions of recordings for which no topic code was assigned as well as segments to which more than one topic code was applied. Overall, the coding schema, training, reliability standards, and procedures were the same for direct audio coding and transcription coding. The methods differed on how codes were applied within the qualitative software project – either to the audio file or to the time-stamped transcript.

3.1.3. Direct audio coding

Data assistants completed direct audio coding using the qualitative software. Once assigned a recorded session, data assistants listened to the audio recording in the software program. As they listened, they made note of the time that discussion of the topic began and ended. Then, they paused the audio file and applied the code to the identified segment. This process was repeated for the entire recording.

3.1.4. Transcription coding

Procedures were also developed for applying codes to transcripts. After a recording was transcribed, it was assigned to a data assistant who read and coded the transcript in the qualitative software. Codes were applied to relevant, timestamped lines of each transcript. Thus, time spent on specific topics was consistently measured across transcription and direct audio coding.

3.1.5. Analysis

Inter-rater agreement and reliability were calculated for each activity code by comparing the codes assigned with each method (direct audio coding or transcript coding), using time as the unit of analysis. Agreement was measured in two ways, both of which were calculated by the qualitative software program: (a) Cohen's kappa, and (b) total agreement. Total agreement was defined as percentage of content, measured by time, coded by both raters and neither rater. This allowed for assessing agreement in a way that accounted for chance agreement between the two raters. The values of the kappa statistic range from zero (random agreement) to one (perfect agreement; Cohen, 1960), and can be used to assess the strength of agreement between raters (Hallgren, 2012; Landis & Koch, 1977). These standards indicate that *K* values above .41 are described as moderate agreement (.41–.60), substantial agreement (.61–.80), and almost perfect agreement (.81–1.0; Landis & Koch, 1977). In instances of complete agreement (100 %) between raters, *K* was not calculated, because chance agreement could not be calculated and accounted for and was usually the result of both raters not applying a code throughout an entire recorded session. For example, very few recordings included the substance abuse code. As a result, both raters were often in 100 % agreement for not applying the code to any segments of the session.

3.2. Results

The presence, number of references, and inter-rater agreement (measured with both percent agreement and Kappa) were assessed for each of the four activity codes across all 15 audio files in the sample (see

Table 1 Comparison of presence, frequency, and agreement in audio and transcription coding.

Code*	Code presence			Frequency of coding references				Range in inter-rater agreement
				Audio		Transcription		
	Audio Only	Transcription Only	Both	*n*	%	*n*	%	
Behavioral and Mental Health	0	0	15	1307	82.3 %	1122	81.7 %	92.07 % –98.58 %
Child Education	1	0	12	211	13.3 %	187	13.6 %	97.27 % –99.97 %
Physical Healthcare	1	1	7	60	3.8%	56	4.1%	98.55 % – 99.99 %
Substance Abuse	0	0	1	11	0.7%	8	0.6 %	99.44 %

* n=15

Table 1). Only one code (Behavioral and Mental Health) was applied in all 15 sessions analyzed. This code was applied most frequently by both methods, with 1,307 references (82.3 % of all references) in audio coding and 1,122 references (81.7 % of all references) in transcription coding. While other codes were applied less frequently, all codes were applied in at least one recorded session. Inter-rater agreement, as measured by percent agreement, was greater than 90 % across all codes, ranging from 92.07% to 99.99%.

Inter-rater reliability was also measured between the raters, each of whom were applying codes to a different type of file (audio or transcription), through calculation of the Kappa statistic (see **Table 2**). Kappa was only measured for sessions where the code was found to be present by both coders. The *Behavioral and Mental Health* code, the most frequently applied code, agreement was substantial to almost perfect. While few references were made to the *Child Education* code in audio (13.3 %) and transcription (13.6 %) coding, inter-rater agreement was substantial (9.09 %) or almost perfect (81.82 %) for 90.9 % of the 11 sessions in which this code was applied, and fair for an additional 9.09 % (n = 1). The *Substance Abuse* code was only applied in one recorded session, however agreement was almost perfect (K = .925). Finally, the *Physical*

Table 2 Proportion of Comparison Study Sessions by Code and Kappa Value.

Code	*Total Recorded Sessions with Code (N)*
Behavioral and Mental Health	15
Child Education	11
Physical Healthcare	7
Substance Abuse	1

Health code was applied by both raters in seven sessions, and inter-rater reliability was substantial (28.57 %) or almost perfect (42.86%) for five sessions and was fair for the remaining two (28.57 %).

3.3. Discussion

In the comparison study, the methods of direct audio coding and transcription coding were compared. Both methods identified the presence and frequency of codes at similar rates **(**e.g., the largest difference in coding frequency across all codes was 0.6 % for Behavior and Mental Health). The percent agreement between raters was greater than 90 % for all codes applied in all recorded sessions. Furthermore, the Kappa coefficient measured substantial or almost perfect agreement across all codes and recorded sessions, except in three instances.

While only small levels of disagreement were measured, it is difficult to know if this resulted from the use of different coding methods, or if it is due to difference in interpretation that would exist between coders using the same method (e.g., both transcription or both audio coding). Alternatively, it could be that the format played a role in the coding of the topic. Perhaps there is something different about hearing the conversation with natural pauses or seeing the words on paper that influenced how raters coded the content. It should also be noted that the most frequently applied code (behavioral and mental health) had high levels of reliability across all fifteen recorded sessions. Additional research is needed to explore if agreement and reliability rates would change for other, less frequently used codes if they were applied with similar frequency.

Overall, the purpose of this comparison study was to better understand how the results of direct audio coding compared to the results of transcription coding. Findings indicate that direct audio coding produced very similar results to transcription coding. This was not only in terms of presence of codes and frequency of application across recorded sessions, but raters achieved high levels of agreement when comparing sessions coded by both methods (> 90 % across all codes). Furthermore, for the most frequently applied codes, reliability measures indicate substantial to almost perfect agreement. The results of the comparison study, therefore, indicated direct audio coding may serve as an appropriate alternate to transcription coding.

4. Reliability study

The direct audio coding method was applied to the project's larger in-home family intervention program evaluation to monitor implementation of service delivery. This study was designed to determine whether data assistants could reliably apply core program specific codes using the direct audio coding method.

4.1. Method

For the larger research study (Duppong Hurley et al., 2019), 241 recorded sessions were collected and direct audio coded to monitor and report fidelity to the service delivery model. A random subset of 102 recordings (42 %) were selected and coded by two raters to assess inter-rater reliability. The setting and data collection procedures of this study were as described in the general study methods.

4.1.1. Coding procedures

The direct audio coding procedures implemented in the evaluation study were similar to those implemented in the pilot study (see Comparison Study Direct Audio Coding). The same team of data assistants completed coding in both studies, but the procedures differed in four ways. First, this reliability study only implemented direct audio coding because results from the comparison study indicated that outcomes would be similar to those generated by traditional transcription coding. Second, the sample used in the evaluation study (n = 102) included a random selection of all recorded sessions collected for the project that were then coded by two raters to test reliability. Third, because the evaluation study was focused on implementation, the set of codes used was specific to the core components of the intervention, rather than the specific topic codes used in the comparison study (e.g. child behavior/mental health, physical health, etc.). The set of core component codes (n = 8) was also larger than the set of topic codes (n = 4) and included: (a) assessment activities, (b) engagement-relationship building activities, (c) family risk screen and safety activities, (d) parenting skills, (e) service planning and documentation, (f) social network mapping, (g) providing supports and resources, and (h) teaching skills surrounding supports and resources.

The core components of the program should always be present in meetings between the Family Consultant and participant. Therefore, all session segments should have a core component code applied. This is unlike topic codes, which were applied only when specific topics were discussed. Finally, because of this, direct audio coding procedures established that core component codes were applied to any audio segment that was at least 15 s long. Segments less than 15 s duration where a core component was discussed were coded with the preceding or subsequent segment. This procedure ensured that the code was only applied when the core component was focus of service delivery, rather than mentioned briefly (e.g., when a participant and family consultant are discussing parenting skills and the participant asks when they will next fill out a specific assessment, but then the conversation immediately goes back to parenting skills).

4.1.2. Analysis

Inter-rater reliability was assessed with measured agreement (percent of agreement) and Cohen's kappa (K) as in the comparison study. The threshold for acceptable inter-rater reliability was 80 % agreement for each code. When agreement fell below this threshold, the two raters met to discuss and resolve differences.

4.2. Results

Inter-rater agreement across all codes was 97.7 %. While this varied by code (see **Table 3**), agreement was at or above 90 % for all codes (*n*= 8). Kappa statistics indicated agreement between raters was moderate, substantial, or almost perfect for 86.7%–100.0% of recorded sessions, depending on code (see **Table 4**). For the three codes most frequently used in direct audio coding (engagement-relationship building activities, parenting skills, supports and resources) over 90 % of recorded sessions measured agreement that was moderate, substantial, or almost perfect.

Table 3 Evaluation Study Inter-rater Agreement in Direct Audio Coding by Code.

Code	*Agreement*
Engagement-Relationship Building Activities	94.4 %
Family Risk Screen and Safety Activities	99.8 %
Social Network Map	99.9 %
Assessment Activities	99.2 %
Parenting Skills	90.8 %
Teaching Skills Surrounding Supports & Resources	99.0 %
Supports and Resources	96.0 %
Service Planning and Documentation	98.9 %

Table 4 Summary of Evaluation Study Kappa Statistics by Code for 102 Recorded Sessions.

Code	n	K				
		Slight Agreement (≤.20)	Fair Agreement (.21–.40)	Moderate Agreement (.41–.60)	Substantial Agreement (.61–.80)	Almost Perfect Agreement (.81–1)
Engagement-Relationship Building Activities	89	1.12%	7.87 %	11.24 %	44.94 %	34.83 %
Family Risk Screen and Safety Activities	9	0.00 %	0.00 %	44.44 %	22.22 %	33.33 %
Social Network Map	3	0.00 %	0.00 %	0.00 %	0.00 %	100.00 %
Assessment Activities	18	0.00 %	0.00 %	5.56 %	11.11 %	83.33 %
Parenting Skills	98	1.02%	7.14 %	7.14 %	43.88 %	40.82 %
Teaching Skills Surrounding Supports & Resources	15	0.00 %	13.33 %	13.33 %	53.33 %	20.00 %
Providing Supports and Resources	61	3.28%	4.92 %	13.11 %	39.34 %	39.34 %

4.3. Discussion

In the study, the reliability of direct audio coding was tested in implementation monitoring. Results indicate that data assistants were reliable, and Kappa coefficients demonstrated high levels of agreement across codes. Inter-rater agreement was greater than 90 % for all codes, including the most frequently used codes (i.e., engagement-relationship building activities, parenting skills, and providing supports and resources). Variance in agreement likely occurred because of the precision with which codes must be applied in the qualitative software. The software system uses an approach to measure a unit of time that is the media equivalent of a single character of text (Baszeley & Jackson, 2014). As a result, failure of coders to start and end codes at the exact same time led

to measured disagreement, even if the switches occurred within a few seconds. Therefore, measuring agreement during transitions between topics was highly sensitive.

Despite the impact such sensitivities may have had in measuring reliability, the results of this study demonstrates the efficiency of the direct audio coding method for the purpose of thematic analysis in a number of ways. First, coupling direct audio coding and a qualitative data analysis software program allowed for more precise coding, tailoring segments to the exact moment core components started and stopped. This is a contrast to other direct audio coding methods found in the literature, which applied codes to fixed segment lengths (e.g., 3 min; Neal et al., 2015). Second, the use of direct audio coding benefited the intervention's fidelity assessment because the results provided more detailed information about the frequency and length of discussions specific to each code. For example, codes could be compared according to their presence in each recorded session, as well as the total amount of time they were discussed in each audio recording. These totals were then be summarized for the entire project and reported in the program evaluation, and proved to be important in the overall fidelity monitoring. Third, the use of qualitative software allowed data assistants to revisit and listen to segments of the observations, by theme when needed, just as one could reread a transcript. While audio files cannot be searched for specific text like a transcript, use of audio allowed data assistants to hear details, such as pausing and tone of voice (Crichton & Childs, 2005). These details could inform coding and were unavailable in the transcript. Overall, direct audio coding with qualitative software provided a number of advantages that outweigh the benefits of a transcript, within the context of implementation monitoring in program evaluation. While this study did not measure time and cost savings, future research should account for these variables to better compare the methods and understand the advantages of the direct audio coding method.

5. Lessons learned

Overall, we found that raters were able to code service delivery sessions reliably between direct audio coding and transcription coding. Moreover, in an evaluation context, we found high levels of inter-rater

reliability when using direct audio coding to assess core intervention components related to the implementation. In this way, the use of direct audio coding with qualitative software may provide a viable approach when transcription is not feasible due to time and cost restraints. Furthermore, the procedures we developed and implemented specific to direct audio coding were effective and supported the overall project evaluation with timely implementation data. This included providing training about the method in a way that allowed multiple data assistants to become reliable. Direct audio coding was then used to analyze a large data set quickly. While the method did not require as much time as transcription, it yielded similar results in terms of inter-rater agreement and reliability.

Throughout this project, our team learned a great deal about the benefits and challenges of using a qualitative analysis software program. The use of this software benefited our studies in a number of ways. In the comparison study, the software allowed for importing recorded sessions, transcribing recordings, and coding both, which ultimately allowed for the comparison of the two methods. The software also allowed us to create a project, or file, which contained all recorded sessions. This was then saved to a sever where it could be easily accessed by all research team members. Additionally, the qualitative software allowed for quickly aggregating results across a large sample of recordings.

While the software offered advantages, the team also encountered challenges when using it. First, the software program was complex and required intensive training for each member of the data team. In the future, costs associated with this training should be included in analysis of the savings provided by direct audio coding when compared to transcription coding. It should be noted, however, that this initial software training was a one time cost, because data assistants could then use the skills they developed in other projects using the qualitative software. Second, while it is clear that the qualitative software precisely measures agreement, it is not clear how reliability and agreement scores are influenced by this precision. The software did offer an option to "code near" when running reports of agreement and reliability, however this feature was not used because it is unknown exactly how "near" the coded segments need to begin and end to measure agreement. It would have been helpful for the software to offer the option for users to adjust this setting to a specific length of time (e.g., .25s, .5s). Finally, the software provided

two ways in which inter-rater agreement could be assessed: percent of agreement or calculation of the kappa statistic. However, there are multiple other methods (e.g., Gwet's AC_1, Krippendorff's alpha, the Brennan-Prediger coefficient) by which to test agreement (Gwet, 2016). It would be helpful for future versions of the qualitative software to offer users options regarding how agreement is assessed, but such options were not available at the time of this study.

6. Implications

These two studies make unique contributions to the literature in a number of ways. Coding qualitative data without transcripts has been used in similar evaluation contexts (Greenwood et al., 2017; Neal et al., 2015; Skillman et al., 2018). However, this study is unique in its use of direct audio coding to monitor fidelity of service delivery. The data collection method used in this study (observation) also differs from those used in other applications of coding without transcription in the literature, including interviews (Neal et al., 2015), focus groups (Greenwood et al., 2017; Mosavel, Ferrell, & Gokee LaRose, 2018), or both (Skillman et al., 2018). The comparison study added to the limited research which compares transcription coding and direct audio coding (Greenwood et al., 2017), while helping to demonstrate that direct audio coding yields similar results to transcription coding. The method was then applied in the reliability study to 102 observations, a sample far greater than the number of records analyzed in previous studies (e.g., Neal et al., 2015; Greenwood et al., 2017). A final distinguishing characteristic of this study was the use of qualitative data analysis software to directly code recorded sessions, as opposed to listening to the audio recordings and taking notes (Greenwood et al., 2017) or using a coding form (Neal et al., 2015).

While direct audio coding has direct implications for researchers and evaluators, there are also implications for practitioners. In the reliability study, direct audio coding was used in implementation monitoring as part of the larger evaluation plan. However, for service providers, routine implementation checks are important to quality service delivery and outcomes, and may be conducted within or outside of a formal evaluation. Such monitoring, though, can be costly and time-consuming,

especially if observations are transcribed. Thus, the use of direct audio coding to monitor fidelity could make qualitative data collection and analysis more feasible for practitioners, allowing for quick feedback that can inform course-corrections related to quality of service delivery for program managers and staff.

Limitations of direct audio coding may be related to both the purpose and context of this study. In our studies, direct audio coding was conducted in a research lab by university data assistants. The lab had access not only to qualitative data analysis software, but also had the time and resources to provide training, supervision, and to check reliability. This method was used to provide timely feedback specific to implementation fidelity within the context of a program evaluation. As a result, and as noted by Neal et al. (2015), use of methods like direct audio coding may not be best suited in different research contexts or with other theoretical foundations and methodologies (e.g., ethnography). However, researchers and practitioners may benefit from continuing to explore the use of direct-audio coding in implementation monitoring and in other evaluation settings where timely and cost-effective feedback is of paramount importance. This includes additional research which compares the reliability of direct audio coding and transcription coding, which would expand understanding of the method's utility and build upon limited comparisons in the literature (Greenwood et al., 2017). Finally, at the conclusion of the project we became interested in how, specifically, direct audio coding may have provided the project with time and cost savings, especially given the high costs of transcription documented in the literature (Kvale & Brinkmann, 2009; Neal et al., 2015; Skillman et al., 2018; Tessier, 2012; Tracy, 2013). While direct audio coding eliminates costs associated with transcription, future research should incorporate measures of time and cost savings in order to best assess any benefits associated with direct audio coding.

7. Conclusion

These studies were unique in their testing and application of direct audio coding, which was found to have results consistent with transcription coding and high rates of inter-rater agreement and reliability. This contributes to the limited literature in which the method of direct audio

coding is used in a program evaluation context. Results demonstrate that direct audio coding has utility in monitoring implementation in service delivery. By maximizing advances in technology available through qualitative data analysis software, direct audio coding allowed for quick and reliable coding of core program elements without a substantial loss of quality. While additional research is needed to continue to explore the utility and validity of direct audio coding, this method is likely to benefit others with similar constraints regarding the time and cost of qualitative data coding.

Funding The development and preparation of this article was supported in part by a research contract Father Flanagan's Boys Town and a training grant from the Institute of Education Sciences (IES), U.S. Department of Education [#R324B160033]. The opinions expressed are those of the authors and do not represent views of Father Flannagan's Boys Town, the Institute of Education Sciences, or the U.S. Department of Education.

Competing Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution

Jennifer Farley: Formal analysis, Writing original draft, Writing review & editing, Visualization, Methodology.

Kristin Duppong Hurley: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing review & editing.

A. Angelique Aitken: Writing original draft, Writing review & editing.

Acknowledgments The authors wish to thank Jay Ringle for coordinating the audio data collection, Lori Synhorst for leading the coding training and reliability efforts, and all the students and families that were a part of the study.

References

Basit, T. (2003). Manual or electronic? The role of coding in qualitative data analysis. *Educational Research, 45*, 143–154. Baszeley, P., & Jackson, K. (2014). *Qualitative data analysis with NVIVO.* Thousand Oaks, CA: SAGE.

Christie, C. A., & Fleischer, D. N. (2010). Insight into evaluation practice: A content analysis of designs and methods used in evaluation studies published in North American Evaluation-Focused Journals. *The American Journal of Evaluation, 31*(3) 326-246.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cope, D. G. (2014). Computer-assisted qualitative data analysis software. *Oncology Nursing Forum, 41*, 322–323.

Crichton, S., & Childs, E. (2005). Clipping and coding audio files: A research method to enable participant voice. *International Journal of Qualitative Methods, 4*(3), 2–9.

Crichton, S., & Kinash, S. (2003). Virtual ethnography: Interactive interviewing online as method. *Canadian Journal of Learning and Technology/La revue canadienne de l'ap- prentissage et de la technologie, 29*(2).

Duppong Hurley, Kristin, Lambert, Matthew, Patwardhen, Irina, Ringle, Jay, Thompson, Ron, & Farley, Jennifer (2019). Parental report of outcomes from a randomized trial of in-home family services. *Journal of Family Psychology, 34*(1), 79–89. https://doi.org/10.1037/fam0000594

Greenwood, M., Kendrick, T., Davies, H., & Gill, F. J. (2017). Hearing voices: Comparing two methods for analysis of focus group data. *Applied Nursing Research, 35*, 90–93. Gwet, Kilem (2016). Testing the difference of correlated agreement coefficients for statistical significance. *Educational and Psychological Measurement, 76*(4), 609–637. https://doi.org/10.1177/0013164415596420

Halcomb, E. J., & Davidson, P. M. (2006). Is verbatim transcription of interview data always necessary? *Applied Nursing Research, 19*, 38–42.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*, 23–34.

Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the craft of qualitative research interviewing.* Thousand Oaks, CA: SAGE.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Leech, N. L., & Onwuegbuzie, A. J. (2007). An array of qualitative data analysis tools: A call for data analysis triangulation. *School Psychology Quarterly, 22*, 557–584.

Mosavel, M., Ferrell, D., & Gokee LaRose, J. (2018). *House chats as a grassroots engagement methodology in community-based participatory research: The WE project*, *10*, Petersburg: Progress in Community Health Partnerships, 391–400.

Neal, J. W., Neal, Z. P., VanDyke, E., & Kornbluh, M. (2015). Expediting the analysis of qualitative data in evaluation: A procedure for the Rapid Identification of Themes From Audio Recordings (RITA). *The American Journal of Evaluation, 36*, 118–132.

Poland, B. (1995). Transcription quality as an aspect of rigor in qualitative research. *Qualitative Inquiry, 1*(3), 290–310. QSR International (2016). *NVivo 11* [software]. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Saldaña, J. (2016). *The coding manual for qualitative researchers.* London, England: SAGE. Skillman, M., Cross-Barnet, C., Friedman Singer, R., Rotondo, C., Ruiz, S., & Moiduddin, A. (2018). A framework for rigorous qualitative research as a component of mixed method rapid-cycle evaluation. *Qualitative Health Research, 29*, 279–289.

Tessier, S. (2012). From field notes, to transcripts, to tape recordings: Evolution or combination? *International Journal of Qualitative Methods, 11*(4), 446–460.

Tracy, S. J. (2013). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact.* Malden, MA: Wiley-Blackwell.

Vander Putten, J., & Nolen, A. (2010). Comparing results from constant comparative and computer software methods: A reflection about qualitative data analysis. *Journal of Ethnographic and Qualitative Research, 5*, 99–112.

Woods, M., Paulus, T., Atkins, D. P., & Macklin, R. (2016). Advancing qualitative research using Qualitative Data Analysis Software (QDAS)? Reviewing potential versus practice in published studies using ATLAS.ti and NVivo, 1994-2013. *Social Science Computer Review, 34*, 597–617.

Jennifer Farley is an IES Postdoctoral Research Fellow in the Academy for Child and Family Well Being at the University of Nebraska – Lincoln. Her research focuses on interventions that promote parent engagement and support teachers and administrators to build a positive school culture and climate. She is currently working to identify supports for parents of students receiving special education services, including students with emotional and behavioral challenges, and analyze how parental involvement is measured.

Kristin Duppong Hurley is a research professor in the Department of Special Education and Communication Disorders, and the co-director of the Academy for Child and Family Well Being at the University of Nebraska-Lincoln. Her focus is on services research for youth with emotional and behavioral needs. Currently she is directing research to improve parental engagement in their child's school and mental health services through parent-to-parent phone support. Dr. Duppong Hurley is also evaluating in-home services to improve parenting and family-functioning with at-risk families.

Angelique Aitken is an Institute of Education Sciences Postdoctoral Research Fellow in the Academy of Child and Family Well-Being at the University of Nebraska-Lincoln. Her scholarship addresses literacy instruction, specifically for struggling writers and the educators who support them, in the general and special education contexts. Within this field, she has two interconnected lines of inquiry: writing intervention and writing motivation. To answer her research questions she employs quantitative, qualitative, and mixed methods methodologies.