Connecting dots: Coding multiple data sources to enhance qualitative analysis

Abstract

This research paper elaborates on the process used by a team of researchers to create a codebook from interviews of Civil Engineers, which included students, professors, and professionals, solving ill-structured problems. The participants solved two ill-structured problems while speaking aloud their thought process. In addition to recording the participant verbalization, the solution to their problems were also collected with the use of a smart pen. Creating a codebook from interviews is a key element of qualitative analysis forming the basis for coding. While individuals can create codebooks for analysis, a team-based approach is advantageous especially when dealing with large amounts of data. A team-based approach involves an iterative process of inter-rater reliability essential to the trustworthiness of the data obtained by coding. In addition to coding the transcripts as a team, which consisted of novice, intermediate, and experts in the engineering education field, the audio and written solution to the problems were also coded. The use of multiple data sources to obtain data, and not just the verbatim transcripts, is lesser studied in engineering education literature and provides opportunities for a more detailed qualitative analysis.

Initial codes were created from existing literature, which were refined through an iterative process. This process consisted of coding data, team consensus on coded data, codebook refinement, and recoding data with the refined codes. Results show that coding verbatim transcripts might not provide an accurate representation of the problem-solving processes participants used to solve the ill-structured problem. Benefits, challenges and recommendations regarding the use of multiple sources to obtain data are discussed while considering the amount of time required to conduct such analysis.

Introduction

Coding verbatim transcripts is general practice in Engineering Education

Qualitative research provides richer data as it gives a deeper understanding of the research question beyond what quantitative data can provide. For example, qualitative inquiry can provide intricate details about why students drop out of the engineering field [1]. Qualitative methods in engineering education can be used as a primary or secondary method. It is becoming popular as demonstrated by the increase in its use in past 15 years [2] and the push for its quality in the engineering education research [3]. Qualitative data usually involves the use of interview transcripts or open-ended questions which are analyzed by coders using a codebook. Coding can be done by a single coder or a team of coders. A team-based approach to coding qualitative data allows for processing of larger amounts of data. Qualitative analysis is a time-consuming process and heavily relies on inter-rater reliability for trustworthiness.

Difference between coding verbatim transcript, linked audio and field notes

Qualitative data is usually collected and analyzed in the form of text transcripts, field notes, and audio/video recordings. While literature on how to conduct qualitative research is abundant (e.g.

[4] - [7]), they lack a discussion on handling audio data. Each data source differs in terms of the level of detail they provide in coding, ability to capture intonation, coding time, and the software requirements [8]. For example, coding verbatim transcripts allows for detailed coding as compared to the audio recordings, but results in loss of nonverbal information which cannot be captured by transcripts [9]. Data sources such as audio recordings can be linked to verbatim transcripts with the help of software to overcome the loss of nonverbal data. Qualitative analysis is enhanced by combining such data sources to achieve higher levels of detail in coding by capturing verbal and nonverbal information than either of the data sources analyzed individually [10].

Coding audio in Engineering Education research

Verbal protocol analysis is one example where audio is used as a measure to plot and compare the difference in use of various problem-solving process used by participants to solve ill-structured problems [11] - [13]. Verbal protocol involves participants talking out loud about what they are thinking while they solve or perform a task [11]. Atman and Burisic [11] used the amount of text coded on transcripts as proportional to the time spent in various problem-solving processes. Others (e.g. [11], [12]) coded the time by looking at videos of participants solving the problem and segmenting it according to various problem-solving processes. While previous research has used time as a metric, there is limited discussion in the literature on the appropriate handling of such data.

Prior to coding these data sources, one must establish if there is indeed a difference between the results obtained from the different data sources. Whether the additional time required to code using videos translates to benefits, since qualitative analysis is time consuming. To explore such aspects of the data, this study had the following research questions:

- Is there a difference in data obtained by coding verbatim transcripts and coding multiple data sources?
- What benefits and challenges exist when obtaining data by coding multiple data sources?
- What practices are recommended for implementing and analyzing multiple data sources?

Methods

Participants

The participants consisted of 7 students, 5 professors, and 4 professionals. Each participant was asked to solve two ill-structured problems while speaking their thoughts out loud i.e. verbalization. The participants were asked to solve simple engineering problems prior to solving the ill-structured problems, so that they were familiar with the process of verbalization. Each participant had 35 minutes to solve the ill-structured problem which included both reading the problem and solving it.

Data collection

To record the solutions, participants were provided with a smart pen which recorded the audio and the solution as it was written in real time. This recording not only recorded the solution but what the participant was verbalizing while writing their solution. Finally, the real time solution with the audio was exported to an interactive video file (i.e. field notes, see Figure 1). One can click on any

part of the solution in the file provides the audio and written solution in real time. The audio recordings from the smart pen were transcribed by a professional transcriber with timestamps.

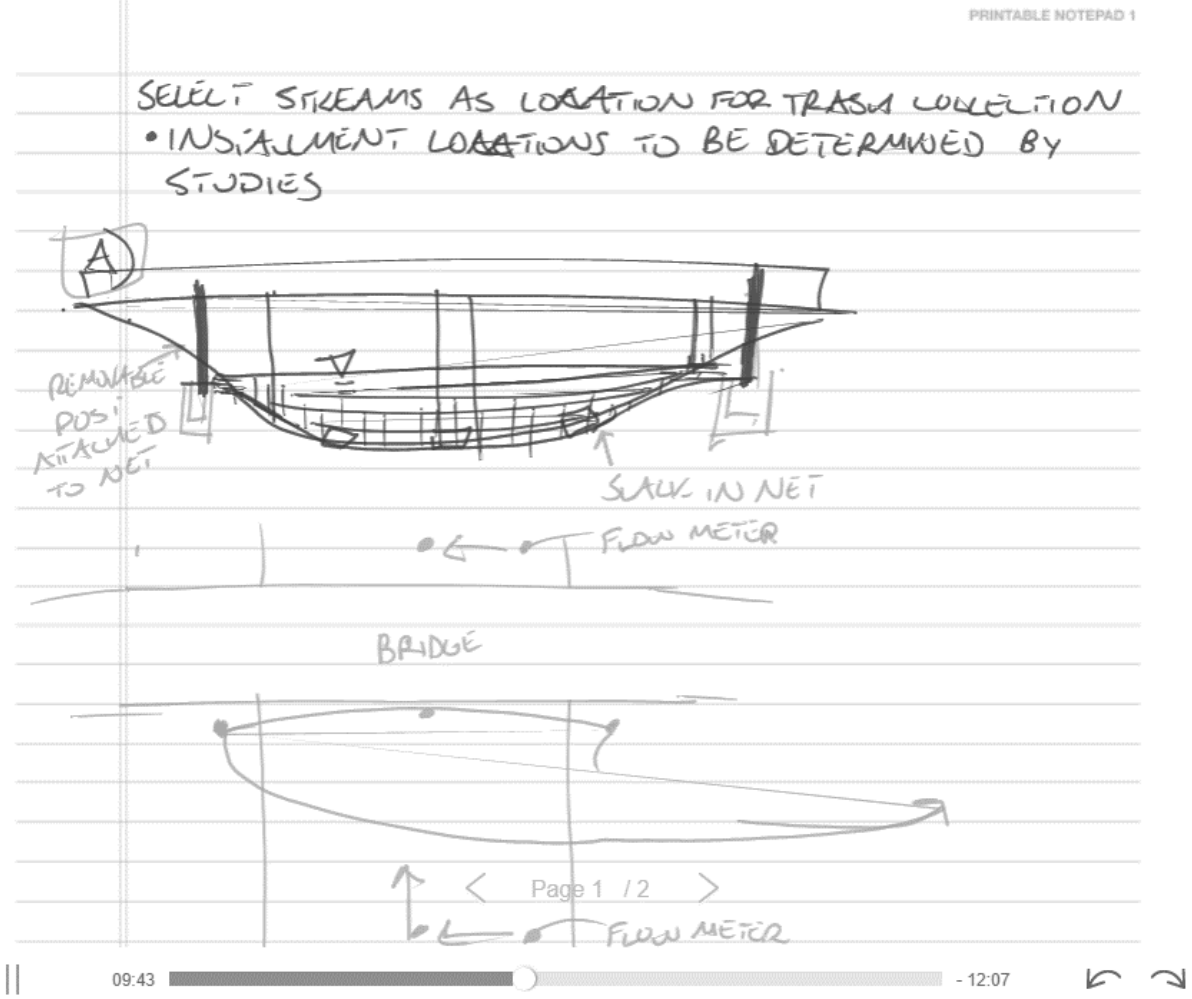

Figure 1. Example screenshot from the interactive video file which shows the participant making notes in real time

Coding transcripts collaboratively

Projects which involve teams of researcher advocate that qualitative analysis should be a collaborative effort [7], [14]. The first step to qualitative analysis in this study was to create an initial codebook from the existing literature. Codes from previous research were collected in a codebook which contained the names of the code, description or definition of the code, and examples of the code. Coding was done by a team of four researchers which included both graduate and undergraduate students. The two graduate students were assigned the role of "lead" coders, one student kept notes of the meeting discussion and agenda for the next meeting, and one student kept a track of the discussion time for each agenda. The role of the lead coders was to modify, refine, and update the codebook according to the meeting discussions. The initial codebook was created prior to viewing any of the transcripts and contained coding instructions for coders. Once the research team was familiar with the codebook each team member coded the same transcript. The research team then met weekly to compare their codes and refine the codebook until a

consensus was reached on the codes and definitions (Intercoder Agreement 1; see Figure 2). This round of coding also served as training sessions for the novice researchers in qualitative coding. Lead coders also conducted tutorials in the use of software and qualitative analysis during these sessions. The coded transcripts were then recoded from the meeting discussions by the lead coders. The duration of each of the inter coder meetings was an hour and half, which the research team deemed to be adequate. This initial coding done by the research team on one transcript was the first round of coding.

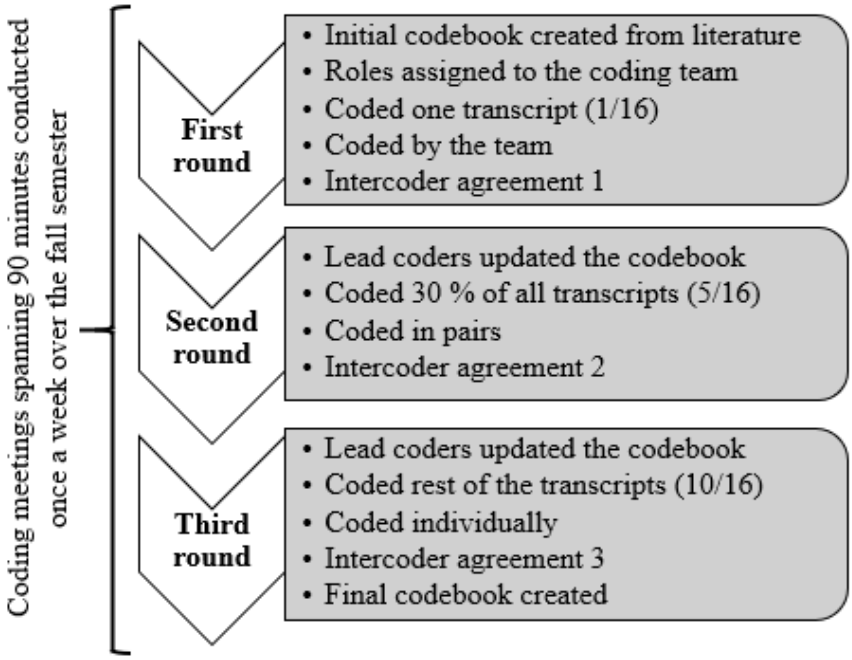

Figure 2. Description of the coding rounds for all transcripts

The second round of coding started with the lead coders creating a refined codebook from the first round. Five of the sixteen transcripts (30 %) were randomly selected to capture all variations in the responses [15]. Each of the five transcripts were coded by a team of two coders. Each coder pair met prior to the team coding meeting to reach a consensus on their coded transcripts. All discrepancies arising in the transcripts were resolved during the team coding meetings, where input from the full research team helped reach consensus (Intercoder Agreement 2; see Figure 2). Each of the five transcripts were coded in this manner and a refined codebook was created which contained the definition, description of when to use, when not to use, and examples as suggested by MacQueen et al. [16].

The third round of coding involved coding rest of the transcripts individually. Coders were assigned transcripts and any issues that arose during the individual coding were brought up at the research team meetings where it was resolved based on input from the entire research team (Intercoder Agreement 3; see Figure 2). Throughout the rounds of the coding process, the research team relied on intensive discussions and group consensus to reach intercoder agreement rather than using qualitative measures such as the Kappa coefficient since this entire process is interpretive [7]. Each intercoder agreement, though slightly different from the others, involved refining, adding, and removing codes that were predetermined from literature to be aware of the subtleties unique to this research [17].

Linking multiple data sources (transcript, audio, and field notes)

In addition to collaborative coding, the transcripts, audio files and field notes were also linked together using the following process:

1. Link the transcripts with the audio with the help of the coding software, MAXQDA. This software recognizes the timestamps on the transcripts and automatically links the audio and the verbatim transcript.
2. Note the start and stop times of when a section of the interactive video file is active. Sections were chosen to represent the participant's complete train of thought.
3. Screenshot the section of the interactive video file associated with a part of a transcript and paste in a word document with timestamp (start and stop) noted.
4. Import audio, transcript, and word document into the coding software.
5. Listen to audio file according to the timestamps in the word document to make sure they are correct.
6. Map codes the audio file codes according to the coded transcript and the word document.

Data Analysis

To conduct data analysis for purpose of comparison of methods, several definitions are needed. First, *real coverage* is the percentage of the transcript covered by the coded segment. *Normalized coverage* is the percentage of the coded problem covered by the coded segment. The *text coverage* is obtained by coding the transcripts only. The *time coverage* is obtained by coding the audio files through the process of linking all the data sources (transcript, audio, and field notes) available. Two transcripts (4 questions) were randomly selected from the second round of coding and linked using the process described above and analyzed. To see if there was a difference in the text and time coverage for each coded segment a paired sample t-test was employed. To compare the two coverages obtained from coding verbatim transcripts and multiple sources of each coded segment, this study employed two-tail paired sample t-tests and set the alpha level at 0.05.

Results and Discussion

Is there a difference?

The comparison in Table 1 shows that coding multiple data sources allows for intricate coding because it captures more information about the interviews and participants than coding verbatim transcripts. Neal et al. [8] provides a similar comparison for coding approaches using each data source individually including coding verbatim transcripts. The quantitative research questions that are explored from the data collected from coding transcripts usually use frequency counting [18] while coding audio allows for comparing frequency count and time taken by each problem-solving process [13].

One could argue that the amount of text is proportional to the amount of time spent in a problem-solving process. This information is easily calculated from coding verbatim transcripts and does not require the coding of additional data sources (i.e. audio files and field notes). The coverage data is even calculated automatically by the coding software in many cases. While the proportion of text to audio maybe true in some cases, e.g. the second question for participant 2 (P2:Q2), the paired sample t-test (Table 2) for the coded segments for the two questions (Q1, Q2) for each

participant (P1, P2) shows that coding the transcript and coding audio give different results. Half of the questions analyzed for difference in text and audio coverage show statistically significant difference i.e. P1:Q1 and P2:Q2. One of the questions showed coverage difference between the two, but it was not statistically significant according to the set alpha level (0.05).

Table 1. Difference between coding transcripts and coding multiple data sources

	Traditional coding of verbatim transcripts	**Coding of transcripts, linked audio file and field notes**
Level of detail	*High.* Intricate coding is possible but results in loss of nonverbal information	*Very high.* Intricate coding is possible with integrated audio and field notes preserving verbal and nonverbal information
Ability to capture intonation	*Low.* Difficult to capture in text. Usually done by providing description of intonation in brackets between text of the transcription.	*Very high.* Intonation is preserved by the audio in addition to data on time taken by the participant engaged in the problem-solving process from field notes.
Coding time	*Slow.* 360 min to transcribe 60 min audio plus time for coding the transcript	*Very slow.* 360 min to transcribe 60 min audio plus time for coding the transcript. Coding audio files and field notes takes approximately an additional 50 % time over coding verbatim transcripts.
Software required	*Optional.* Required if transcription is done by researchers. Not necessarily required for just coding text.	*Required.* Transcripts, audio files, and field notes are linked by software. e.g. MAXQDA
Quantitative research questions	*Yes*. Quantitative research questions can use statistical techniques based on frequency count only (e.g. [18]).	*Yes.* Quantitative research questions using statistical techniques based on frequency count and time (e.g. [13]).

Table 2. Mean (%), variance, and t-test results for real coverage for text and time for each question

Question	**Text**		**Audio**		**t-test**	
	Mean	**Var**	**Mean**	**Var**	**df**	**Sig.**
P1: Q1	0.43	0.61	0.57	1.32	65	**
P1: Q2	0.41	0.21	0.46	0.33	32	+
P2: Q1	1.31	6.16	4.34	68.30	22	*
P2: Q2	1.10	3.54	1.11	6.51	34	

Note: +p<0.1, *p<0.05, **p<0.01, ***p<0.001

In addition to capturing more information and hence giving richer data about the problem-solving process, coding multiple data sources provides more accurate information on the proportion of problem-solving process used by participants to solve the problem.

What are the benefits and challenges?

The goal of studying students, professors, and professionals ill-structured problem-solving skills was to see if any group was inclined to use certain process more than the others. Processes such as *idea generation, idea expansion, idea comparison, feasibility assessment, hypothetical process, and using outside knowledge* were identified from previous literature and refined through the iterative coding process in the codebook.

Most qualitative analysis software calculates the real text coverage obtained from coding verbatim transcripts. From Table 1 we see that the information provided differs from the real time coverage obtained from multiple sources. This implies that coding only the text-based transcripts have bias associated with them. For example, if a participant is inclined and comfortable with the verbalization process, they will provide a richer transcript than a person who is not comfortable with the verbalization process. The latter may have spent the same or more time on certain processes as compared to the former which was missed in coding the text only transcripts. Certain process coverages are prone to being under-captured or over-captured by the text coding. This may be consistent for everyone because one can never expect perfect verbalization. Hence, using such data for quantitative analysis apart from simple counting leads to erroneous results.

To gather data regarding problem solving processes, this study linked multiple data sources to obtain the time coverage of each of the processes. While the process of linking these data sources was fairly straightforward, and only required additional time to link and map, there were some instances where special attention was required. Instances where participants verbalized everything they wrote with the smart pen were simple to code, map, and did not show any difference in coverage (Table 2 P2:Q2). Instances when the participant did not verbalize but still made notes with the smart pen or verbalized something different while they made notes with the smart pen (the most common was interactions with the interviewer while writing the solutions) required changing the coded segments to match the ill-structured problem-solving process accurately. This difference in mapping of the audio files from the transcript is shown by the coverage difference (Table 2, P1:Q1, P2:Q2).

Recommendation

While extracting accurate data should be the goal of every research team, the time payoff for coding the additional data source takes, in the case of this effort, as much as 50 % more time than that required to code the verbatim transcripts. This can have significant impacts on the timeline of the research project.

Table 3. Mean (%), variance and t-test results for the normalized coverage for text and time for each question

Question	Text		Audio		t-test	
	Mean	Var	Mean	Var	df	Sig.
P1: Q1	1.52	7.56	1.52	9.28	65	
P1: Q2	3.03	11.38	3.03	14.41	32	
P2: Q1	4.43	60.52	4.34	68.30	22	
P2: Q2	2.86	23.74	2.86	42.87	34	

Note: +p<0.1, *p<0.05, **p<0.01, ***p<0.001

To overcome this, the data was normalized. Normalized coverage was calculated as a percentage of the coded transcript, not the complete transcript. The difference between coded and complete transcript is that the coded transcript coverage does not included parts of the transcripts that were not coded (e.g. interactions with interviewers or unrelated thoughts). Table 3 shows that by normalizing the two coverages the difference between the two data sets (i.e. text and audio) is removed. Normalization of the transcript coverages could be a good measure to compare the use of ill structured problem-solving processes while not having to code the additional data sources.

Conclusion and Implications

The results indicate that the data resulting from coding the verbatim transcripts do not match the actual amount of time taken in most cases. In some cases the differences are statistically significant and in other cases they are not. Thus analysis using verbatim transcripts should be used with awareness of its limitations. Using such data to compare the relative amount of time associated with different problem solving processes between students, professors, and professionals has some implications on the results. The level of impact depends on the how comfortable the participant under consideration is at verbalization.

For some processes it is beneficial to use multiple data sources to calculate coverages. This provides the most accurate representation of the processes. The time required to code these additional data sources, however, cannot be predicted which may be challenging to manage to meet a research timeline. To balance the time payoff and accurate representation of the processes, this study suggests normalizing the coverage data which reduces the bias at a statistical level. Most qualitative analysis software provides coverages which can be easily normalized without the requirement of additional time as compared to coding multiple data sources. Engineering education is seeing a growth of qualitative studies; thus, it has become more important to assess the potential of qualitative methods and revisit strategies that are used to conduct such analysis. The study will continue to code more transcripts in the above methodology to improve and inform qualitative analysis in engineering education.

Funding

This work was funded by National Science Foundation Grant DUE #1712195. The project is entitled "*Collaborative Research: Bridging the gap between academia and industry in approaches for solving ill-structured problems*". Data, findings, and conclusions or recommendations are those of the authors, only.

References

1. M. Meyer and N. Fang, "A qualitative case study of persistence of engineering undergraduates," *International Journal of Engineering Education*, vol. 35, 1, pp. 99-108, 2019.
2. E. Douglas, "Beyond the interpretive: Finding meaning in qualitative data," in *Proceedings of the 124th Annual Conference and Exposition: American Society for Engineering Education, Columbus, OH, USA, June 25-28, 2017*.
3. J. Walther, N. Sochacka, L. Benson, A. Bumbaco, N. Kellam, A. Pawley, and C. Phillips, "Qualitative research quality: A collaborative inquiry across multiple methodological perspectives," *Journal of Engineering Education,* vol. 106, 3, pp. 398-430, 2017.

4. J. Maxwell, *Qualitative Research Design: An Interactive Approach*. Thousand Oaks, CA.: SAGE Publications, 2013.
5. L. Richards, *Handling Qualitative Data: A Practical Guide* Thousand Oaks, CA.: SAGE Publications, 2015.
6. D. Hays and A. Singh, *Qualitative Inquiry in Clinical and Educational settings*. New York, NY: Guilford Press, 2012.
7. J. Saldaña, *The Coding Manual for Qualitative Researchers*. Thousand Oaks, CA.: SAGE Publications, 2013.
8. J. Neal, Z. Neal, E. VanDyke and M. Kornbluh, "Expediting the analysis of qualitative data in evaluation: A procedure for the rapid identification of themes from audio recordings (RITA)," *American Journal of Evaluation*, vol. 36, 1, pp. 118–132, 2015.
9. S. Tessier, "From field notes, to transcripts, to tape recordings: Evolution or combination?" *International Journal of Qualitative Methods*, vol. 11, 446–460, 2012.
10. J. Evers, "From the past into the future. How technological developments change our ways of data collection, transcription, and analysis," *Qualitative Social Research*, vol. 12, 1, pp. 1–31, 2011.
11. C. Atman and K. Bursic, "Verbal protocol analysis as a method to document engineering student design processes." *Journal of Engineering Education*, vol. 87, 2, pp. 121-32, 1998.
12. C. Atman, M. Cardella, J. Turns, and R. Adams. "Comparing freshman and senior engineering design processes: An in-depth follow-up study." *Design Studies,* vol. 26, 4, pp. 325-57, 2005.
13. C. Atman, R. Adams, M. Cardella, J. Turns, S. Mosborg, and J. Saleem, "Engineering design processes: A comparison of students and expert practitioners," *Journal of Engineering Education*, vol. 96, 4, pp. 359-379, 2007.
14. M. Schreier, *Qualitative content analysis in practice*. Thousand Oaks, CA.: SAGE Publications, 2012.
15. D. Hruschka, D. Schwartz, D. St. John, E. Picone-Decaro, R. Jenkins and J. Carey, "Reliability in coding open-ended data: Lessons learned from HIV behavioral research," *Field Methods*, vol. 16, 3, pp. 307-331, 2004.
16. K. MacQueen, E. McLellan, K. Kay and B. Milstein, "Codebook development for team-based qualitative analysis," *Cultural Anthropology Methods*, vol. 10, 2, pp 31-36, 1998.
17. G. Spindler and L. Spindler, "Cultural process and ethnography: An anthropological perspective," In *The handbook of qualitative research in education,* M. D. LeCompte, W. L. Millroy, and J. Preissle, San Diego: Academic Press, 1992, pp. 53-92.
18. R. Dixon, A. Raymond, and S. Johnson, "Experts vs. novices: Differences in how mental representations are used in engineering design" *Journal of Technology Education*, vol. 23, 1, pp. 47-65, 2011.