For the Record: Exploring variability in interpretations of police investigative interviews

Felicity Deamer, Emma Richardson†, Nabanita Basu & Kate Haworth

Aston Institute for Forensic Linguistics, Aston University, UK
& †Loughborough University, UK

https://doi.org/10.21747/21833745/lanlaw/9_1a2

Abstract. *Recent research (Haworth 2018) has demonstrated how investigative interview data are (unintentionally) distorted as they pass through the criminal justice system, and the survey-based experiment we present here was designed to test our hypothesis that various aspects of the processing of police-suspect interview data may have an impact on the quality of the official evidential document produced. The quantitative and qualitative findings from this experiment shed light on, and provide a sound evidence base for this claim, rather than leaving it as an untested assumption. The experiment was designed to test each key aspect of the current process of the production of routine written transcripts of investigative interviews (ROTIs), focusing on the conversion from spoken to written format, and the use of different transcription conventions, and it has enabled us to investigate which changes make the most difference in terms of the evidential quality of the end product, in order to effect a change in practice which will reduce or eliminate the effect of those changes. Our findings suggest that when presented with a transcript of a police interview, we are significantly more likely to (1) perceive the interviewee as anxious and unrelaxed, (2) interpret the interviewee's behaviour as being agitated, aggressive, defensive, and nervous, (3) determine that the interviewee is un-calm and uncooperative, and (4) deem the interviewee's version of events to be untrue, than we are if we listen to the original audio recording. Moreover, subjects identified (a) consistency, (b) phrase and lexical choice, (c) emotion (crying/upset), (d) hesitation and/or pauses as significant factors influencing participants' perception and interpretation of the interviewee and their story. This is particularly concerning as the latter two features are not currently routinely included in police transcripts, and Haworth (2018) illustrates multiple ways in which transcripts might differ from the original audio recordings they are intended to replace, with respect to words and phrases, as well as general content. The findings presented in this paper provide a strong motivation for further research into how we capture spoken interaction in legal contexts, and they constitute something of a mandate for reform with respect to the transcription of police interviews in the UK.*

Keywords: *Investigative interviews, Transcription, Entextualisation, Interpretation, Perception.*

Resumo. *Investigação recente (Haworth 2018) mostrou como os dados das entrevistas policiais são (involuntariamente) distorcidos durante o processamento pelo sistema de justiça criminal. O inquérito experimental que apresentamos neste artigo procura testar a nossa hipótese de que vários aspetos do tratamento dos dados das entrevistas policiais com os suspeitos podem influenciar a qualidade da prova oficial. Os resultados quantitativos e qualitativos desta experiência elucidam-nos e fornecem uma fundamentação sólida para esta assunção, até agora não testada. A experiência foi concebida para testar todos os aspetos centrais do atual processo de produção de transcrições do processo de entrevistas de investigação (ROTIs), concentrando-se na conversão da forma oral para a forma escrita e na utilização de diferentes convenções de transcrição, o que nos permitiu investigar quais as alterações que exercem maior impacto em termos de qualidade probatória do produto final, com o intuito de introduzir alterações na prática e assim reduzir ou eliminar o efeito dessas alterações. As nossas conclusões sugerem que, perante transcrições de uma entrevista com a polícia, é significativamente mais provável (1) percecionar o entrevistado como ansioso e inquieto, (2) interpretar o comportamento do entrevistado como agitado, agressivo, defensivo e nervoso, (3) percecionar o entrevistado como pouco calmo e cooperante, e (4) considerar que a versão dos eventos do entrevistado não é verdadeira, contrariamente ao que acontece se ouvirmos a gravação áudio original. Além disso, os participantes identificaram (a) consistência, (b) seleção frásica e lexical, (c) emoção (choro/perturbação), (d) hesitação e/ou pausas como fatores relevantes que influenciam a perceção e interpretação do entrevistado e da sua história, o que é particularmente preocupante uma vez que as duas últimas características não integram atualmente o processo de transcrições policiais. Como mostra Haworth (2018), as transcrições podem diferir, em diversos aspetos, das gravações áudio originais que substituem, não só em palavras e frases, mas também conteúdo geral. As conclusões apresentadas neste artigo constituem um forte incentivo a mais investigação sobre a interação verbal em contextos legais e constituem um apelo à reforma do processo de transcrição das entrevistas policiais no Reino Unido.*

Palavras-chave: *Investigative interviews, Transcription, Entextualisation, Interpretation, Perception.*

Introduction

Police investigative interviews play a critical role in the criminal justice system: they are one of the primary methods of evidence gathering during an investigation and can later serve as crucial evidence during a trial. Standard procedure in England and Wales is that police interviews are audio recorded, then transcribed by clerks employed by the relevant police force. This process is of particular importance given that these are evidential documents, routinely presented in court as part of the prosecution case, yet the original spoken data are (necessarily) substantially altered through the process of being converted into written format. Once a transcript, typically referred to as a ROTI (Record of Taped Interview), has been produced, it is generally heavily relied upon in place of the audio recording, especially in court, making its accuracy all the more important.

Haworth (2018) argues that interview evidence is unintentionally distorted and misinterpreted as it journeys through the criminal justice system; from an initial interaction that takes place in a police interview room to a transcript of an audio recording of that interview being read out in court (typically by the Prosecution) during a trial. Haworth draws attention to the asymmetry between the loose and unregulated practices and procedures associated with the handling of police interview evidence and the strict principles of preservation applied to physical evidence (e.g., DNA, blood spatter, fingerprint evidence), which she argues results from a lack of recognition that changes in the format of linguistic data involve a transformation of the data themselves. Haworth highlights both accidental and intentional discrepancies between the original interview and the version which takes its place as evidence in court. These include poor audio recording and lack of detail and nuance in transcription conventions (e.g. no pauses, intonation, stress emphasis, emotion, overlapping speech etc.), alongside deliberate editing and summarising to condense the official record. Yet the output of each transformational process is treated as an all but identical copy of the previous version (2018: 434).

It is of course, important to recognise that even without any such shortcomings in practice surrounding audio capture and transcription, it isn't possible to create a perfectly accurate written version of spoken language. There will always be something of a translation process involved in converting spoken language into written. Levelt (1983, 1989), among others, illustrated that listeners have a natural tendency to 'repair' any disfluencies they hear in speech, allowing them to make sense of what is being said. Collins *et al.* (2019) looked at what happens if disfluencies are embraced and incorporated into transcripts, focusing on whether filled pauses are perceived in the same way within transcripts as they are in speech. They found that disfluencies in speech (i.e. fillers such as 'um' and 'er') were more likely to be perceived as indicators of uncertainty in the speaker when presented in text (as part of a transcript) than when heard on an audio recording. This suggests that it is virtually impossible to capture and represent speech in written form without some distortion taking place. Moreover, (Fraser 2003) draws our attention to the fact that almost all attempts to convert spoken language into written form are made using an audio recording rather than the original face-to-face interaction (for obvious practical reasons). This means that however good the audio recording is, the talk in question has already been stripped of its meaningful context, and physically present and animated speakers, thus disarming our perceptual and inferential capacities of that critical information. Add to that the fact that there will inevitably be numerous points in any audio recording in which the sound quality is poor enough to leave the listener in doubt with respect to exactly what they are hearing, and it is clear that transcription necessitates a certain amount of informed guess work. The danger is that as transcribers we tend not to be aware of our perceptual inaccuracies, and as with all perception we are not conscious of the role that our predictions and expectations play in our experience of any given perceptual input. (**?**) refers to this as 'the unacknowledged role of the perceiver' (2003: 204), emphasising 'the active role we play in constructing the messages we hear by combining the information in the speech signal with the knowledge in our heads' (2003: 206; see also (Fraser 2014, 2018). (Bucholtz 2009) emphasises similar considerations with respect to the role of the transcriber in the specific context of the processing of spoken data within the legal system. (Coates and Myths 1999) recognise

the distortions that occur when spoken language is captured and converted into text, and argue that transcripts should be treated as nothing more than "an analytic convenience to make [spoken] data accessible to readers".

There is a pressing need for transcription guidelines and training to assist ROTI transcribers in producing ROTIs which encapsulate more of the meaning conveyed by the original spoken interaction, and to enable consistency of interpretation of features such as punctuation and pauses for the reader (i.e., fellow investigating officers, lawyers, courts). We are therefore currently undertaking a project to (1) test our hypothesis that there is a serious unrecognised problem within the criminal justice system with evidential consistency in investigative interview records; (2) collaboratively develop standard guidance for transcription, using transcription conventions which can easily be incorporated into practice; and (3) design new, linguistically-based input to the training for ROTI transcribers. A substantial increase in the accuracy and standardisation of investigative interview evidence (especially in terms of the representation of spoken language features) would enable those who use ROTIs as evidence to be able to interpret punctuation or other visual representation of spoken features consistently when they occur (regardless of who it was produced by) thus removing a major source of potentially subjective and inaccurate interpretation of criminal evidence. However, the above research points to the fact that any such endeavour must be based on evidence, not intuitions or expectations about how spoken and written data are perceived. Our first step, therefore, was to start building that evidence base. The study we present here is the first (as far as we are aware) to directly compare perceptions of and interpretations drawn from a transcript and an audio recording of the same spoken interaction.

The specific aim of the small-scale experiment discussed here was to assess individual perceptions of different versions of the same interview data (audio vs. written transcript). This is the first in a series of studies designed to test our hypothesis that aspects of the processing of police-suspect interview data have a negative impact on the quality of the official evidential document produced. Across this series of experiments, we aim to test each key aspect of the current process of the production of routine written transcripts of investigative interviews (ROTIs), focusing first in this current experiment on the conversion from spoken to written format, before then moving on to testing (in subsequent studies) the use of different transcription conventions. Our long-term intention is to investigate which changes make the most difference in terms of the evidential quality of the end product, in order to effect a change in practice which will reduce or eliminate (the effect of) those changes.

We emphasise that all versions in which interview interaction are recorded will inevitably involve a degree of alteration of the data, and there is no such thing as a 'perfect' transcript Fraser (2003); we also acknowledge that all use of data as evidence involves subjective interpretation on the part of the judge or jury, and that these interpretations will therefore inevitably vary to some extent. What we are aiming towards through this series of experiments (of which this is the first) is assessing which written version is evidentially closest to the original, in terms of introducing the least amount of change in interpretation when compared to the "purest" version available (here, the audio/video).

Method

In order to accurately compare individual perceptions of different versions of the same police interview, we assessed participants' interpretations, impressions, and judgements of both the interviewee themselves and what was said in the interview. These assessments were carried out under experimental conditions in which the mode of presentation was manipulated (transcript or audio recording) in order to assess the effect on the participants' perception of the interviewee, their interpretation of what was said in the interview, and their overall judgement with respect to the truth or falsity of the interviewee's version of events.

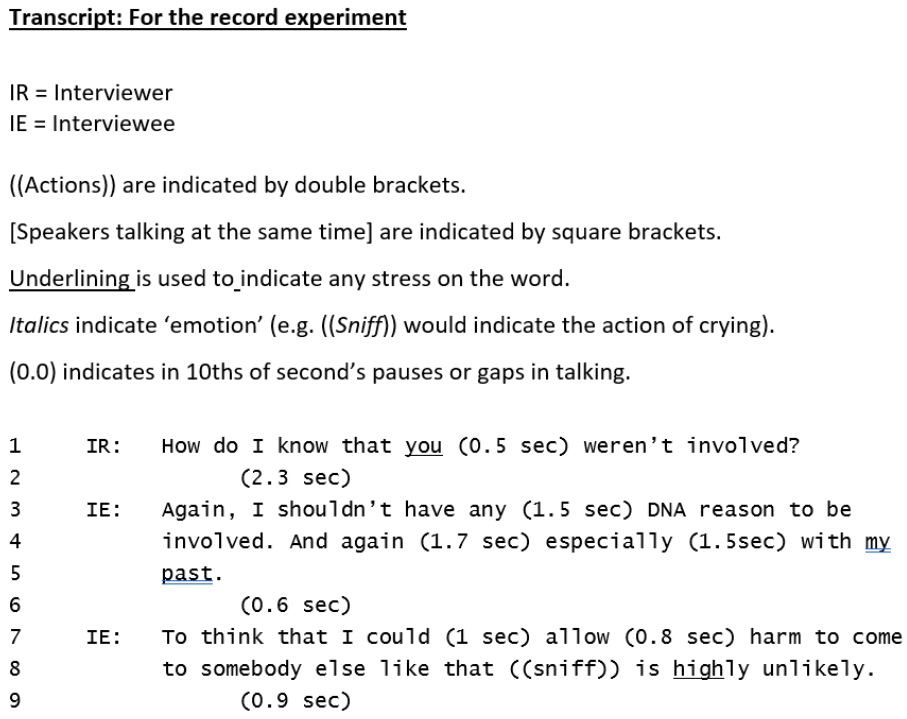
Participants

Sixty adult native speakers of English were invited to take part in the study using convenience sampling. All participants were contacts of the research team, and had been identified as (a) not having any prior linguistic training, and (b) not having any knowledge of the research project and its aims and objectives. Following recruitment, participants were emailed a link to the online survey. On opening the survey, participants were all presented with a 3-minute clip[1] from the same police interview (10 minutes long in full), taken from publicly available footage on YouTube of a suspect interview in a UK murder enquiry K.L.E.E. Photography (2015). Thirty participants heard the 3-minute audio recording of the original interview, and the other thirty saw a written transcript of the same clip. The groups were matched for gender and age (approx. even spread from age 18 – 71).

Materials and procedure

The interview clip in the two conditions was exactly the same, only mode of presentation varied between conditions (i.e., written transcript or audio). Participants were encouraged to listen to or to read the interview as many times as they liked, for as long as they liked prior to and while answering the questions that followed. Immediately following having heard or read the interview, participants were presented with a series of questions, some of which elicited quantitative data (i.e., number on a Likert scale), and some of which required an open answer in a text box (lending themselves to qualitative analysis). Participants were reminded that they could continue to listen to or to read the interview while answering the questions.

The transcript was produced with two key considerations in mind, (1) ensuring legibility for a lay audience, and (2) ensuring that it included as much detail as possible, given the first consideration of legibility. With these considerations in mind, we were able to include pauses, stress emphasis, overlapping speech, and emotion (transcribed as 'sniff'); all features which have long been established within linguistics as conveying substantial amounts of meaning, and which are therefore generally included in even relatively simple linguistic transcripts Jefferson (2004). The intention was to include as much detail as we could in the transcript, so that we could compare a 'best possible' transcript (given the first consideration above) with the original audio recording. If we had produced a transcript that might be considered closer to an officially produced police transcript / ROTI, we would not have been able to ask participants the same questions of the data (i.e., concerning linguistic features, since that information would not have been available to those participants in the Transcript condition).[2] There was a legend at the top of the transcript to explain the meaning of each of the transcription conventions.



Transcript: For the record experiment

IR = Interviewer
IE = Interviewee

((Actions)) are indicated by double brackets.

[Speakers talking at the same time] are indicated by square brackets.

Underlining is used to indicate any stress on the word.

Italics indicate 'emotion' (e.g. ((*Sniff*)) would indicate the action of crying).

(0.0) indicates in 10ths of second's pauses or gaps in talking.

```
1     IR:   How do I know that you (0.5 sec) weren't involved?
2                 (2.3 sec)
3     IE:   Again, I shouldn't have any (1.5 sec) DNA reason to be
4           involved. And again (1.7 sec) especially (1.5sec) with my
5           past.
6                 (0.6 sec)
7     IE:   To think that I could (1 sec) allow (0.8 sec) harm to come
8           to somebody else like that ((sniff)) is highly unlikely.
9                 (0.9 sec)
```

Figure 1. Image of transcript

Below are the questions that participants were presented with after having read or heard the interview extract. Questions 1 and 5 were formulated with the intention of exploring how the interviewee's character might be evaluated differently depending on whether the original audio recording was heard, or whether a transcript of the interview was read. In a related way, questions 3 and 9 probe participants' evaluation of the interviewee's story, and whether those evaluations differ depending on the format in which the interview is presented. Questions 2, 4, 6 and 10, allowing free text responses for the purpose of qualitative analysis, were included in order to glean as much information from participants as possible with respect to how the linguistic dimensions of the interview (i.e., what is said and how it is said) influenced or informed their judgements and evaluations of the interviewee and their version of events. Questions 7 and 8 allow participants to describe the interviewee's emotional and behavioural profile, again enabling us to explore potential differences between groups. The options used in these questions were adapted from (Ekman 1992) universal emotion categorisation system to ensure that we offered participants the opportunity to describe the interviewee's emotional state as accurately and as thoroughly as possible. We used multiple choice for these questions in order to ensure that the responses were tractable. Question 11 was included at the end of the survey with a view to establishing whether the format in which the interview is presented to participants impacts on their overall perception of the interviewee's guilt or lack thereof.

1. On a scale of 1 to 5, please specify your level of agreement with the statement below:

   "The interviewee is credible"

"The interviewee is credible"

	1 - I don't agree at all	2	3	4	5 - I agree entirely
"The interviewee is credible"	☐	☐	☐	☐	☐

2. What is it about the language used and/or how it is said that led you to your conclusion about how credible the interviewee is?
3. On a scale of 1 to 5, please indicate how plausible the interviewee's story is.

	1 - Totally implausible	2	3	4	5 - Highly plausible
The interviewee's story is...	☐	☐	☐	☐	☐

4. What is it about the language used and/or how it is said that led you to your conclusion about whether or not the interviewee's story is plausible?
5. On a scale of 1 to 5, please indicate how sincere the interviewee is.

	1 - Totally insincere	2	3	4	5 - Very sincere
The interviewee is...	☐	☐	☐	☐	☐

6. What is it about the language used and/or how it is said that led you to your conclusion about whether or not the interviewee is sincere?
7. On a scale of 1-5, please indicate to what degree the following words could be used to describe the interviewee's emotions at any point during the interview:

	1 - Not at all	2	3	4	5 - Very much
RELAXED	☐	☐	☐	☐	☐
ANXIOUS	☐	☐	☐	☐	☐
FEARFUL	☐	☐	☐	☐	☐
DISGUSTED	☐	☐	☐	☐	☐
SURPRISED	☐	☐	☐	☐	☐
HAPPY	☐	☐	☐	☐	☐
ANGRY	☐	☐	☐	☐	☐
SAD	☐	☐	☐	☐	☐
CONTEMPT	☐	☐	☐	☐	☐

8. On a scale of 1-5, please indicate to what degree the following words could be used to describe the interviewee's behaviour at any point during the interview:

	1 - Not at all	2	3	4	5 - Very much
AGITATED	☐	☐	☐	☐	☐
CALM	☐	☐	☐	☐	☐
PANICKED	☐	☐	☐	☐	☐
FRIENDLY	☐	☐	☐	☐	☐
COOPERATIVE	☐	☐	☐	☐	☐
AGGRESSIVE	☐	☐	☐	☐	☐
DEFENSIVE	☐	☐	☐	☐	☐
ASSERTIVE	☐	☐	☐	☐	☐
NERVOUS	☐	☐	☐	☐	☐

9. In your opinion, is what the interviewee is saying true?

○ Yes
○ No

10. What is it about the language used and/or how it is said that led you to your conclusion about whether or not what the interviewee is saying is true?
11. In your opinion, and based only on what you have heard in the interview, should the interviewee be found Guilty or Not Guilty of being complicit in the murder?

○ Guilty
○ Not Guilty

Analysis

The survey responses were analysed using a mixture of qualitative and quantitative methods.

For the analysis of questions 1, 3, 5, 7, and 8, we analysed the distribution of data for each question in the Audio condition and the Transcript condition. We did this based on measures of central tendency (such as mean, median, and mode) and measures of dispersion (such as standard deviation, inter quartile range, and range.). To test for normality of distribution of responses for each question of each group, Shapiro Wilk test was used (as the number of data points in each group was less than 2000). Given that the datasets deviated from normality assumption, equality of variance for the groups (using Levene's test) was not calculated. Given that normality assumptions were violated and the data is ordinal, a non-parametric test (Mann Whitney U-test) was used to compare the responses for the two groups.

Questions 9 and 11 were dealt with separately because the response to these questions are categorical/nominal and dichotomous. Given that the data (i.e., responses) are nominal and dichotomous, we cannot expect the data to follow a normal distribution. Again, we considered the responses to be unpaired. In order to test whether the distribution of responses is the same for the Audio and Transcript group we used the Chi square test.

Finally, content analysis and inductive thematic analysis were used to derive a coding frame to analyse free text responses to questions 2, 4, 6, and 10 that enabled us to isolate which linguistic features participants identified as contributing to their interpretation of the interview and/or perception of the suspect being interviewed. The coding frame allowed for a nuanced analysis of the specific qualities of the data collected, and was developed iteratively by the research team who met a number of times to scrutinise the text box data and discuss a working list of codes that captured all the linguistic features mentioned in participants' responses. Two researchers separately coded all 60 responses to the free text questions, and then met to discuss and agree on any instances where the two sets of coding didn't match up. The outcome of this coding process is detailed in Table 1.

Code	Audio	Transcript
Clarity	**13**	**23**
In delivery	10	19
In Language	3	4
Content Choice	**37**	**27**
Lexical choice	7	6
Phrase choice	22	18
Repetition	8	3
Emotion	**86**	**43**
Crying/upset	46	27
Genuine+	22	2
Genuine-	8	12
Shock	7	2
Laughter	3	0
Register	**2**	
Formal -	0	0
Formal +	2	0
Sentence Structure	**6**	**6**
Conditionals	4	3
Non-sequiturs	1	0
Unfinished	1	3
Sequencing	**9**	**11**
Interjection	0	1
Question and Answer match	5	4
Quick to answer	4	6
Other		
Hesitation/pausing	14	19
Consistency	3	12
Pace	2	0
Rehearsal	10	8
Sound Quality	1	2
Stress emphasis	0	6

Table 1. Codes and references per condition

Results

Quantitative

The results suggest (see Figure 2 and Figure 3) that there is no statistically significant difference between the responses to question 1 (Z = -1.434, p=.152) and 3 (Z = -.349, p=.727) in the Audio and the Transcript group, indicating that the format in which participants were presented with the interview did not impact on their perception of the interviewee's credibility or the perceived plausibility of the interviewee's version of events[3].

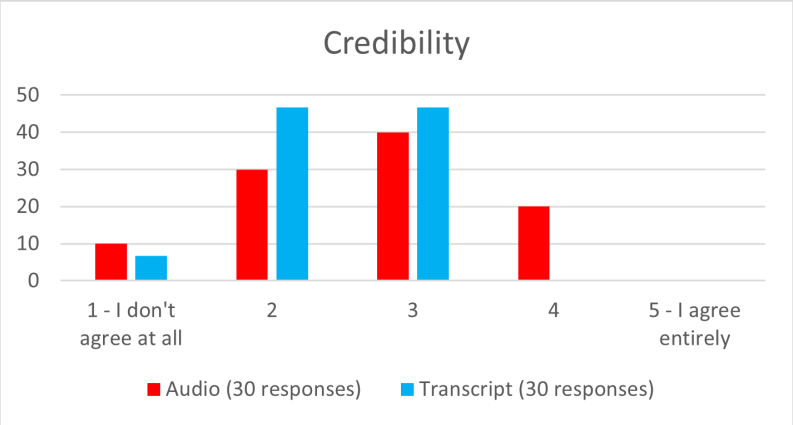


Figure 2. Distribution of responses to question 1 (credibility)

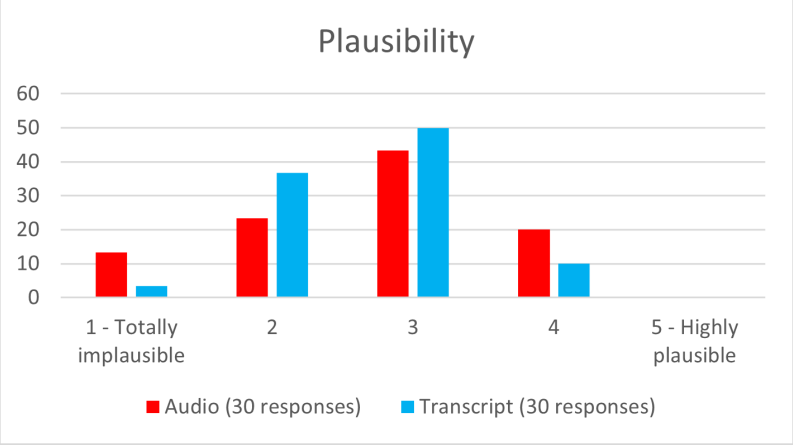


Figure 3. Distribution of responses to question 3 (plausibility)

Results for question 5 (Z=-1.741, p=.082) suggest that with a larger sample size we might find evidence that the format in which participants experience the interview does have an impact on their judgement with respect to how sincere the interviewee is (see figure 4).

The results from question 7 suggest (see Figure 5 to Figure 7) that there is a statistically significant difference (or strong trend towards significance) between the responses in the Audio and the Transcript group, with respect to the attribution of emotional descriptors 'relaxed' (Z= -2.267, p=.023), 'anxious' (Z= -1.984, p=.047), and

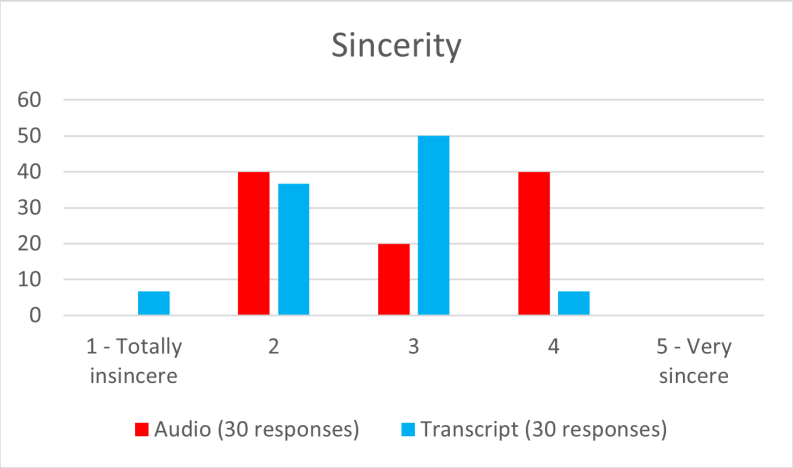


Figure 4. Distribution of responses to question 5 (sincerity)

'fearful' (Z= -1.928, p=.054), suggesting that reading the transcript of the interview makes participants more likely to perceive the interviewee to be unrelaxed, anxious, and fearful than if they were to listen to the original audio recording.

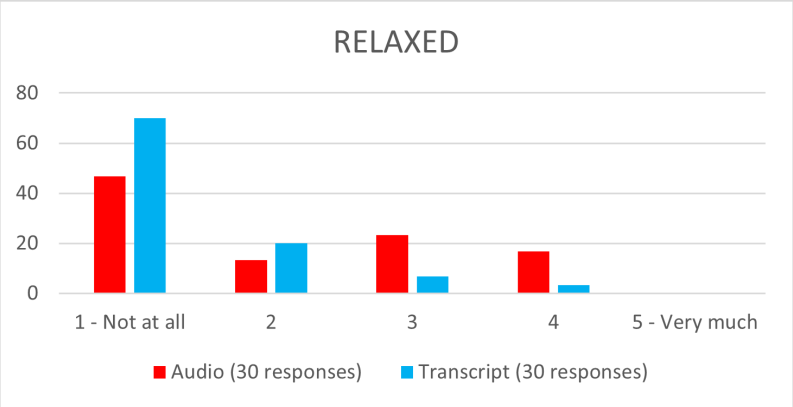


Figure 5. Distribution of responses to question 7 (relaxed)

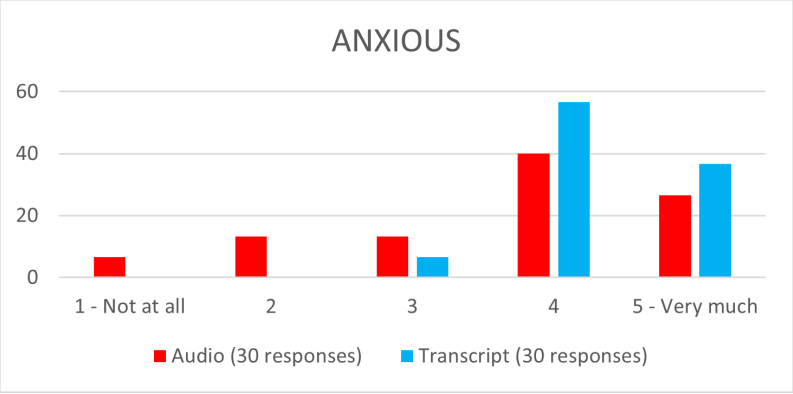


Figure 6. Distribution of responses to question 7 (anxious)

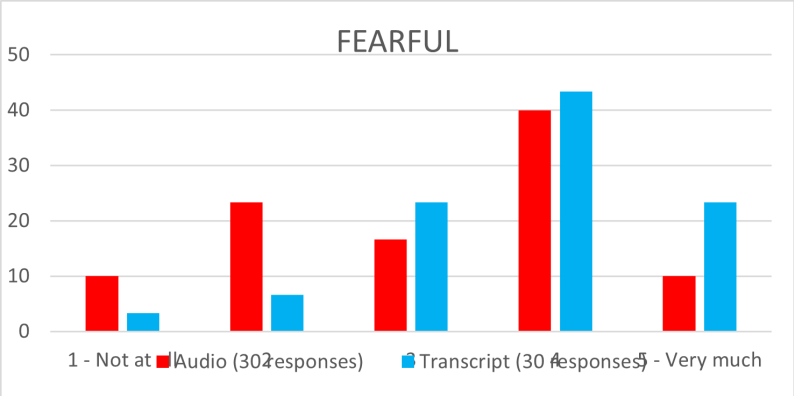


Figure 7. Distribution of responses to question 7 (fearful)

There were no significant differences (see Table 1) between groups with respect to the remaining emotional descriptors 'disgusted' (Z= -.429, p=.668), 'surprised' (Z= -.195, p=.846), 'happy' (Z=-.043, p=.966), 'angry' (Z= -.168, p=.867), 'sad' (Z= -.015, p=.988), 'contempt' (Z= -1.074, p=.283).

Q7_	Relaxed	Anxious	Fearful	Disgusted	Surprised	Happy	Angry	Sad	Contempt
Mann-Whitney U	313.500	326.000	325.500	422.000	437.500	448.500	439.500	449.000	381.000
Wilcoxon W	778.500	791.000	790.500	887.000	902.500	913.500	904.500	914.000	846.000
Z	-2.267	-1.984	-1.928	-.429	-.195	-.043	-.168	-.015	-1.074
Asymp. Sig. (2-tailed)	.668	.846	.966				.867	.988	.283

Table 2. Mann Whitney U Test results for question 7

With regard to Question 8, a further Mann Whitney U test also showed that there was a significant difference in the responses between the Audio and Transcript group with respect to the attribution of behavioural descriptors 'agitated' (Z= -2.95, p=.003), 'calm' (Z= -2.41, p=.016), 'cooperative' (Z= -2.33, p=.020), 'aggressive' (Z= -2.12, p=.034), 'defensive' (Z= -3.32, p=.001), and 'nervous' (Z= -2.39, p=.016) (see figures 8-13), suggesting that reading the transcript of the interview makes participants more likely to perceive the interviewee to be *agitated*, *aggressive*, *defensive*, and *nervous* than if they were to listen to the original audio recording.

There were no significant differences (see Table 2) between groups with respect to the remaining behavioural descriptors 'panicked' (Z= -1.52, p=.128), 'friendly' (Z= -1.74, p=.080), and 'assertive' (Z=-..742, p=.458).

Q8_	Agitated	Calm	Panicked	Friendly	Cooperative	Aggressive	Defensive	Assertive	Nervous
Mann-Whitney U	257.500	298.500	350.000	336.500	302.500	327.500	232.000	402.000	296.000
Wilcoxon W	722.500	763.500	815.000	801.500	767.500	792.500	697.000	867.000	761.000
Z	-2.955	-2.413	-1.522	-1.748	-2.332	-2.121	-3.328	-.742	-2.398
Asymp. Sig. (2-tailed)	.003	.016	.128	.080	.020	.034	.001	.458	.016

Table 3. Mann Whitney U Test results for question 8

A significant difference was recorded (see figure 14 and Table 3) in the distribution of responses to question 9 "In your opinion, is what the interviewee is saying true?"

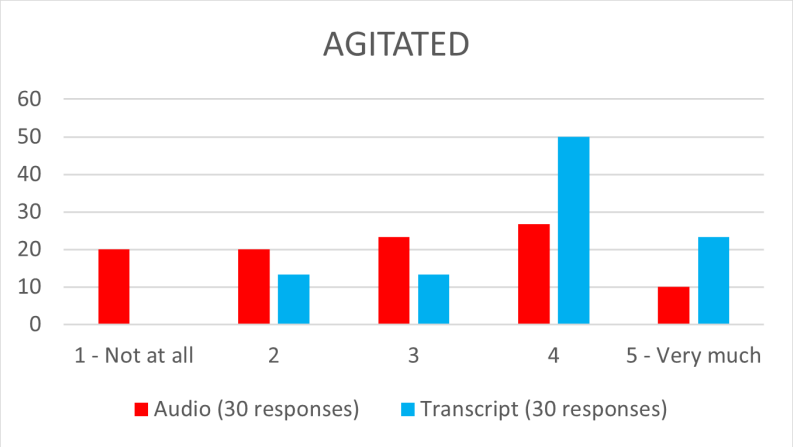


Figure 8. Distribution of responses to question 8 (agitated)

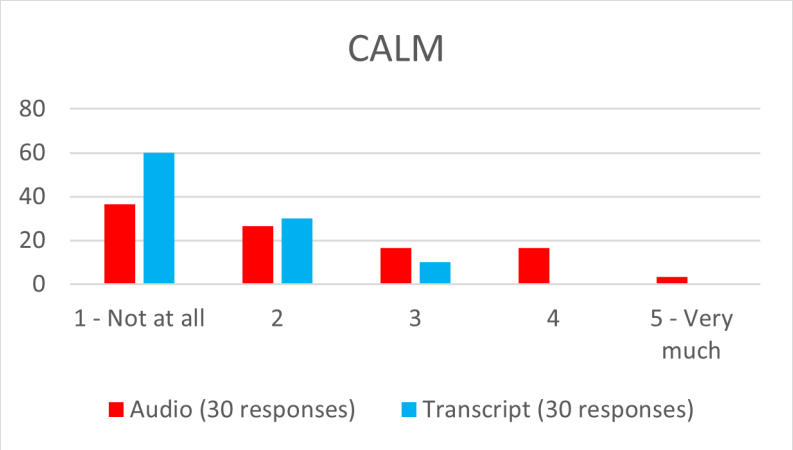


Figure 9. Distribution of responses to question 8 (calm)

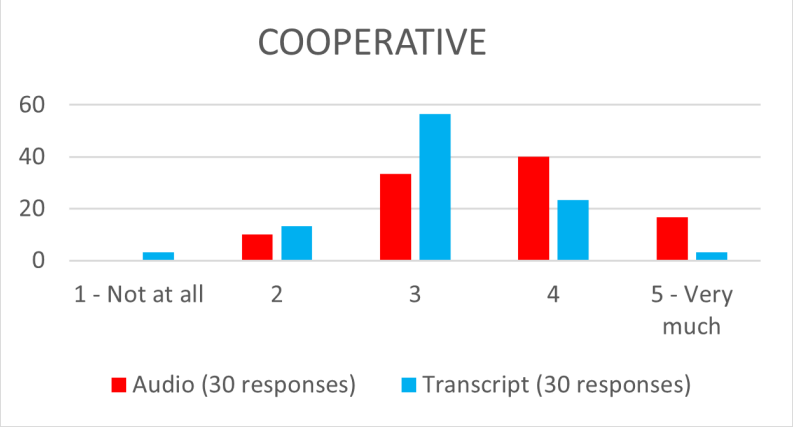


Figure 10. Distribution of responses to question 8 (cooperative)

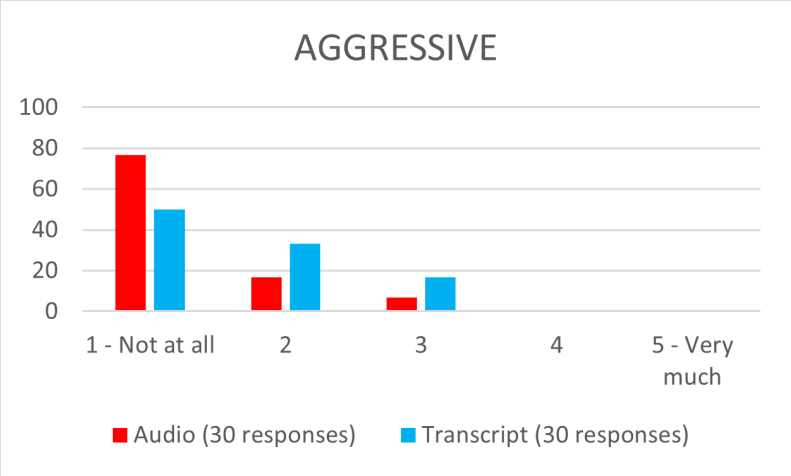


Figure 11. Distribution of responses to question 8 (aggressive)

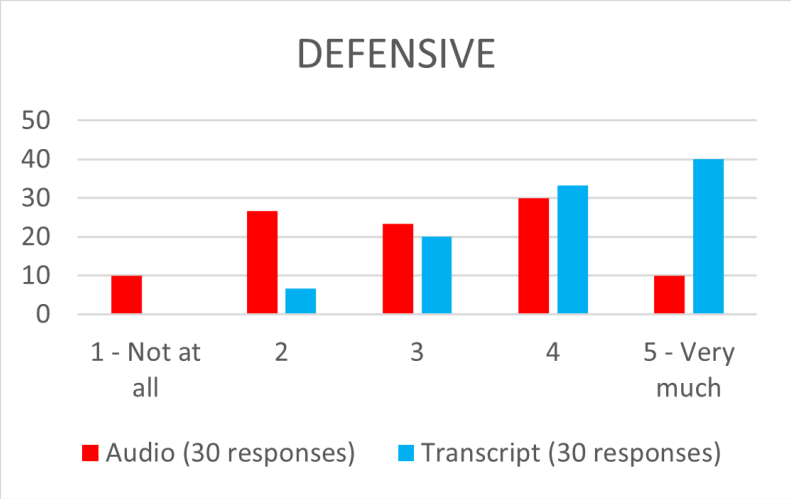


Figure 12. Distribution of responses to question 8 (defensive)

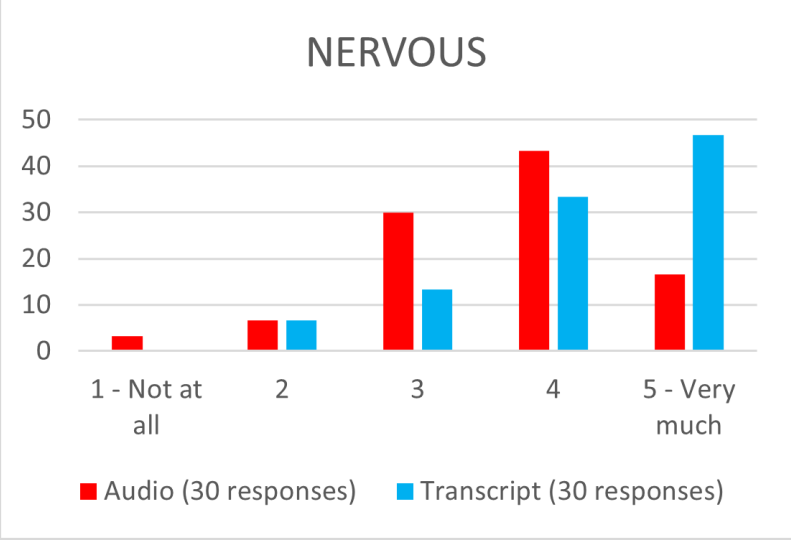


Figure 13. Distribution of responses to question 8 (nervous)

between the Audio and Transcript group ($^2(1) = 4.022$, p=.045), indicating that those who read the transcript of the interview were more likely to judge the interviewee's version of events to be untrue.

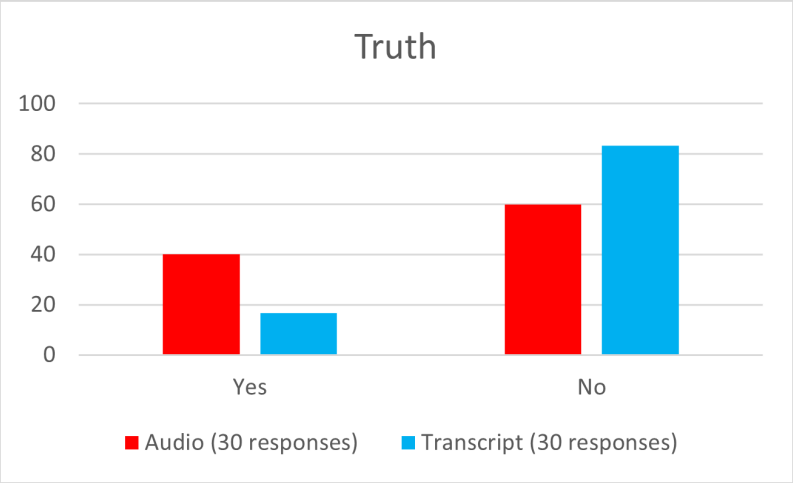


Figure 14. Distribution of responses to question 9 (truth)

No significant difference was recorded in the distribution of responses to question 11 "In your opinion, and based only on what you have heard in the interview, should the interviewee be found Guilty or Not Guilty of being complicit in the murder?" between the Audio and Transcript group ($^2(1) = 0.268$, p=.605).

In summary, participants who read the transcript of the interview were significantly more likely than those who had heard the original audio recording to:

1. Perceive the interviewee as *anxious* and *unrelaxed*
2. Interpret the interviewee's behaviour as being *agitated*, *aggressive*, *defensive*, and *nervous*
3. Determine that the interviewee is *un-calm* and *uncooperative*[4]
4. And ultimately, to deem the interviewee's version of events to be *untrue.*

This study uses a small sample size, and our effect sizes may be a consequence of that. We intend to replicate the findings in a larger study, in which we have (a) a larger sample size in each condition (n=64), and (b) multiple experimental conditions to be compared including *Audio* (the same audio recording used in this current study), *Full Transcript* (the same transcript used in this current study), *ROTI* (a transcript produced in line with standard police practice), *Pauses Removed* (the same as the full transcript condition but with pauses removed), and *Emotion Removed* (the same as full Transcript condition but with markers of emotion removed).

Qualitative

Here, we present the main themes identified from the text responses to the questions: what is it about the language used and/or how it is said that led you to your conclusion about how *credible* the interviewee is (Q2), whether or not the interviewee's story is *plausible (Q4)*, whether or not the interviewee is *sincere (Q6)* and whether or not what the interviewee is saying is *true (Q10).*

Although the quantitative analysis did not find significant difference in responses to the questions of plausibility and credibility, the qualitative analysis enables us to understand further the justifications for scores, based on the language used in the two conditions. We present three main findings, firstly relating to the temporal aspect of the delivery of the suspect's speech, secondly the interpretation of emotion, and finally the impact of an impoverished mode of presentation on participants' judgements of the interviewee and their account.

Temporal aspects of delivery

An analysis of multiple codes ('quick to answer', 'hesitation' and 'rehearsal') outlines how participants reported making use of the speed at which the interviewer's questions were answered by the interviewee to make their judgements. The first of these to be explored here is where the suspect was perceived as being quick to answer the interviewer's questions.

The interviewee was 'Quick to answer'

At the end of the interview extract the interviewer asks the suspect two yes/no interrogative questions, both answered by the suspect with a "no". In the transcript, no pause was indicated prior to the delivery of "no", where pauses are timed and indicated (to the tenth of a second) elsewhere. Participants in each condition reported the speed at which the suspect answers as a justification for their scores.

Extract 1:

> *"The short, sharp "no"s towards the end of the clip seemed insincere." (Audio, Truth No)*

Extract 2:

> *"She was very quick to answer the IR question on line 101." (Transcript, Truth No)*

Extract 3:

> *"She is relatively quick to answer the questions, however doesn't give a lot of detail." (Transcript, Truth Yes)*

Extract 4:

> *"She doesn't hesitate at all - and it backs up the rest of her story that she didn't know what was going on." (Audio, Plausibility 4)*[5]

Most of the references captured by this code relate to this final section of the transcript and whether or not the suspect was telling the truth. If a written record of the investigative interview does not seek to represent the speed at which responses are delivered, then a reader compared with a listener will have unequal opportunity to evaluate the suspect's account in this respect.

'Hesitation' in the suspect's responses

In addition to the speed at which the suspect answered some questions, pauses during the delivery of the suspect's speech (mid-turn pauses) were interpreted by participants as relating to 'thinking' about what to say. Hesitation was used to explain scores mostly on the lower end of the scale for credibility and for why they deemed the suspect not to be telling the truth.

Extract 5:

> *"There were hesitations in her replies as if she was trying to convince herself what she was saying was true as well as trying to convincing the interviewee." (Audio, Credibility 2)*

Extract 6:

> *"Very vague answers, constant hesitation as if they are considering what they say." (Transcript, Credibility 2)*

Extract 7:

> *"[T]here were pauses where it felt like the interviewee was thinking too much about the right thing to say." (Audio, Truth, No)*

However, participants relayed to us that when pauses, interpreted as hesitation, are heard or read in conjunction with 'emotion' in the suspect's speech, this positively impacted their perceptions.

Extract 8:

> *"The nervous, halting, and emotional quality of the delivery make the interviewee sound sincere to me" (Audio Sincerity 4)*

Extract 9:

> *"Not too much hesitation in her voice. Her emotions also come across as genuine, in a way in which I can't articulate." (Audio, Sincerity 4)*

Extract 10:

> *"On the one hand the language, stresses, pauses, could be genuinely indicative of a highly upset individual who is struggling with the circumstances." (Transcript, Credibility 3)*

Whether influencing negatively or positively, these findings argue in favour of including pauses in transcripts, where they feature in the audio, since this demonstrates their importance to the assessment of the interview as evidence.

In addition to pauses, the suspect's use of fillers such as 'uh' and 'um' in their speech also was categorised by one participant as 'hesitation'.

Extract 11:

> *"The constant us[e] of "um" giving herself time to think." (Audio, Credibility 1)*

It is common for these not to be transcribed even in what the police consider to be a "verbatim" transcript. However, the responses from this research indicate that their inclusion could impact on hearers' and readers' interpretation of the suspect's account. The two temporal aspects of delivery discussed so far also feature in the next section where the delivery of the speech is evaluated as if it were, or were not, rehearsed.

The suspect's responses were 'rehearsed'

Participants in both conditions recurrently described the suspect's answers as 'rehearsed', or not. Here we present some of the references which directly attribute that to the temporal aspects of delivery rather than to the content of what is being delivered in the narrative (which we do not focus on in the current paper).

> *"I don't feel the interviewee is sincere as the pauses suggest she has rehearsed what she is going to say, and rather than it being said with true emotion, she has said things for effect." (Transcript, Sincerity 1)*

Extract 12:

> *"In my opinion, what the interviewee is saying sounds rehearsed and unnatural. They don't stutter much, neither do they use "umms" or "urrs", etc. It sounds like something that has been prepared and memorised in order to specifically address the key doubts against them. It seems deliberately vague." (Audio, Truth No)*

Pausing during the speech is described by a participant in extract 13 to be interpreted as planned delivery. In extract 14, we see another participant comment on the lack of disfluency in the response. The responses of participants in relation to plausibility, credibility and sincerity being influenced by a sense of rehearsal does appear to be mitigated again by the presence of what they interpret to be emotion.

Extract 13:

> *"It sounds rehearsed. It sounds like she's reading lines until she starts to cry." (Audio, Sincerity 2)*

Extract 14:

> *"Her emotions appear to be authentic with halting speech, interrupted by crying. She appears not to have pre-rehearsed what she is going to say." (Audio, Sincerity 4)*

In the next section we pay particular attention to the ways in which emotion is perceived and referred to by participants in the two conditions.

Emotion

The code containing the highest number of references in our analysis was 'emotion', with participants making numerous comments about the suspect's emotional state when responding to the questions we asked. This notably included considerations of whether or not the suspect was being genuine, with a number of mentions of the suspect being in shock. Crying seemed particularly marked for participants (mentioned a total of 72 times across both conditions). Interestingly, while participants in both conditions made use of the emotion heard in the suspect's voice, or noted in the transcript, to form judgements about the suspect's account and justify their scores, they did this in slightly different ways. Those in the audio condition interpreted the emotion themselves:

Extract 15:

> *"The interviewee's clear shock and emotional state also makes me believe that she's being honest and is more of a victim in this situation." (Audio, Credibility 4)*

Extract 16:

> *"The interviewee seems most sincere at the beginning of the interview when she is most emotional." (Audio, Sincerity 4)*

Extract 17:

> *"[H]er crying sounded quite real - she seems to be sniffing which is quite hard to make happen." (Audio, Sincerity 2)*

By contrast, those in the Transcript condition described their interpretation as filtered through the transcriber's interpretation and representation of emotion in the transcript.

Extract 18:

> *"The repeated sniffing in the text, and the fluency can be read either way, as genuine emotion, or as an attempt to screen her true feelings." (Transcript, Sincerity 3)*

Extract 19:

> *"I'm told she's emotional and crying at various points." (Transcript, Credibility 3)*

These references to the 'text', or being 'told' about a particular feature, demonstrate that the format distanced participants in the Transcript group from the data in a way which did not seem to occur with the Audio group, by inserting the transcriber in between.

Impoverished mode of presentation

Eight of the 30 participants in the Transcript condition assigned a score of '3', the middle of the scale, for questions relating to interviewee credibility, and plausibility and sincerity of the account. In their text responses they commented that not having access to the audio impacted on their ability to make judgement. They commented how, by reading and not listening, they struggled to be "categorical" in their judgements.

Extract 20:

> *"It's hard to decide, even with the emotion/stress/gaps written down. I feel like I'm still missing information on how she's saying it." (Transcript, Credibility 3)*

More specifically, participants made comment about the difficulty they had in answering the questions based on a reading of rather than a listening to the data. They located the issue as requiring the speaker's 'tone' to make judgements.

Extract 21:

> *"I can't [sic] say without hearing the tone of voice." (Transcript, Sincerity 3)*

Extract 22:

> *"Its [sic] difficult to answer this using the text without hearing tone of voice. "(Transcript, Credibility 3)*

However, references captured by the code 'tone' demonstrate that participants in the Audio group did make judgements based on what they describe as the tone of the suspect's voice.

Extract 23:

> *"I didn't hear anything in the tone of her voice or the words that made me think she was being insincere." (Audio, Sincerity 3)*

Extract 24:

> *"The defensive tone in which she spoke at times in the interview made her testimony appear less sincere." (Audio, Sincerity 3)*

Summary of qualitative analysis

The qualitative analysis makes a strong case for the importance of representing temporal aspects of the speech delivery in the transcript and where possible 'emotion' which (as discussed earlier) we indicated through the transcription of 'sniffs' and where the suspect's voice became altered, or strained, through crying. If pauses, audible sniffs and alterations in delivery were not represented within the transcript (as is typically the case in ROTI transcripts), they would be unavailable to those without access, or the resource to listen to the audio recording. We might hypothesise this to alter the interpretation of the evidence.

Discussion

It is perhaps intuitive that vitally important information is lost in the transformation of spoken interaction into written format. It is somewhat less intuitive that this lost information might have a negative impact on the evidential value of the spoken interaction, but this is what we are seeing evidence of in our data. In this experiment, the aim was to move from the *whether* to the *why*; if we know the mechanisms by which evidential value is negatively impacted, then we can take steps to mitigate that impact.

We might think that in each condition there is a different kind of input (audio vs. written), which (due to their inherent differences) leads to different

inferences/conclusions being drawn. But the idea that there is a stable notion of input is problematic; the inputs themselves will be moulded and shaped by our prior experience (see, for example, increasingly popular Bayesian approaches to cognition Griffiths *et al.* (2008)). Neither the audio recording nor the transcript are stable inputs on which rational inferences operate; top-down information (hypotheses/expectations/prior experience) will determine how we perceive the interview, and will influence the conclusions we draw about the person being interviewed. You could view a transcript as a further distortion of an original interview than the audio recording, having removed all the myriad low level contextual information. However, the reader will instinctively make sense of what they are presented with; they won't simply "suspend judgement" on information that is lacking; they will "fill in". This hypothesis is supported by our findings that participants in the Transcript condition were more, not less, likely to attribute emotional properties to the interviewee, like 'anxious', or 'agitated', even though direct evidence of this is lacking. Put generally, this filling in plays a bigger role in the Transcript condition than in the Audio condition as it is informationally impoverished.

The relevant point here is that the transcript is not just leaving information out, but encouraging a process of filling in**.** The reader doesn't simply fail to see what's there, but might infer things that are not. These "top-down" filling in effects don't only impact rational judgement; they reach down into how the input is experienced at the most basic level[6]. This means that with all the good will (and education/training) in the world, a reader won't be able to avoid the effects of their "priors". Another concern is that these "priors" will vary across participants, meaning that, the more there is this filling in, the less likely responses are to be consistent across participants. In other words, not only would we expect to see less *accuracy* in informationally impoverished conditions (i.e. transcripts), but we will also see more variability, namely, less *consistency* in interpretation.

In adding more detail to our Transcript condition (relative to a standard ROTI), such as emotion, pauses, stress emphasis, it could be argued that we added very explicit 'filling-in' prompts. Just as (Collins *et al.* 2019) found that disfluencies such fillers were more marked in transcripts than in audio recordings, perhaps the detail in our transcription made salient those linguistic features which we typically might not attend to when hearing audio. This raises important questions around whether more transcription detail is in fact better, or whether it makes artificially salient those features which might otherwise be experienced as a 'normal' and unmarked feature of spoken language.

This concern is in fact largely not borne out in our qualitative analysis. It is clear from Table 4 that with respect to pauses and emotion, both features which were detailed in our transcript, there are more mentions in the audio condition than in the Transcript condition, and there are almost no mentions of interjections/overlapping speech in either condition. There were only mentions of stress emphasis in the Transcript condition, albeit only six, so this could indeed be a feature which was made artificially salient when marked in the transcript.

Our qualitative analysis illustrates that the speed of delivery, including pauses and hesitation and representations of emotion were key features that participants identified as contributing to their perception and overall judgement of the interviewee and their

version of events, which suggests that these are aspects of the audio recordings of interviews that we should be looking to retain within standard police transcripts (ROTIs).

The findings discussed here are born out of a small-scale study, and it would be prudent to attempt to replicate these findings on a larger scale, comparing multiple conditions:

Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6
Original audio recording	Standard police transcript (ROTI)	Detailed transcript, including pauses, emotion, and overlapping speech	Condition 3 transcript but with pauses removed	Condition 3 transcript but with overlapping speech removed	Condition 3 transcript but with emotion removed

Table 4. Follow-up experimental conditions

However, these initial findings strongly suggest that the onus is on us as linguists to provide the relevant evidence and training to ensure that official police transcripts (ROTIs) contain as much of the vitally important information contained in the original interview as possible, while being mindful of the context and purpose of transcription, as well as the limited resources available to the police Richardson *et al.* (2022), and being conscious of the 'added salience' effects that transcription detail can have.

Acknwoledgements

The research reported in this paper was funded and conducted at Aston Institute for Forensic Linguistics.

Notes

[1]A subsection of the interview was selected to ensure that participants had a manageable amount of information to attend to in close detail when answering the subsequent questions. The specific 3-minute clip was chosen on the basis that it contained the interviewee's main description of events surrounding the killing.

[2]We will run comparisons with an officially produced transcript in later iterations of this project.

[3]We used significance level (i.e,) of 0.05 for all statistical tests. Sample effect is considered to be statistically significant if p value is less than or equal to the chosen value.

[4]As 'calm' and 'agitated' are antonyms, we expected participants who marked high for agitated would mark low for calm. Including these terms allowed us to check for consistency in participants' answers.

[5]The number detailed here refers to where on the sliding scale (1 (not at all)-5 (very much)) participants rated the interviewee on that particular question

[6]See sine-wave speech, Hollow Mask Illusion, McGurk effect etc.

References

Bucholtz, M. (2009). Captured on tape: Professional hearing and competing entextualizations in the criminal justice system. *Text and Talk*, 29(5), 503–523.

Coates, J. and Myths, J. (1999). Lies and audiotapes: Some thoughts on data transcripts. *Discourse Society*, 10(4), 594–597.

Collins, H., Leonard-Clarke, W. and O'Mahoney, H. (2019). Um, er': how meaning varies between speech and its typed transcript. *Qualitative Research*, 19(6), 653–668.

Ekman, P. (1992). An argument for basic emotions. *Cogn. Emot*, 6, 169–200.

Fraser, H. (2003). Issues in transcription: Factors affecting the reliability of transcripts as evidence in legal cases. *Forensic Linguistics*, 10(2), 203–226.

Fraser, H. (2014). Transcription of indistinct forensic recordings: Problems and solutions from the perspective of phonetic science. *Language and Law/Linguagem e Direito*, 1, 5–21.

Fraser, H. (2018). Forensic transcription: How confident false beliefs about language and speech threaten the right to a fair trial in australia. *Australian Journal of Linguistics*, 50, 129–139.

Griffiths, T., Kemp, C. and Tenenbaum, J. (2008). Bayesian models of cognition. In R. Sun, Ed., *The Cambridge handbook of computational psychology*. Cambridge University Press, 59–100.

Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *The International Journal of Evidence Proof*, 22(4), 428–450.

Jefferson, G. (2004). *Glossary of transcript symbols with an introduction*. Amsterdam/Philadelphia: John Benjamins.

K.L.E.E. Photography, (2015). Becky Watts murder - 3rd police interview (long version). https://www.youtube.com/watch?v=hN4fYTXObcg.

Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition*, 14, 41–104.

Levelt, W. (1989). *Speaking: From intention to articulation*. The MIT Press.

Richardson, E., Haworth, K. and Deamer, F. (2022). For the record: Questioning transcription practices in legal contexts'. *Applied Linguistics*. https://doi.org/10.1093/applin/amac005. Published online 8th Feb 2022.