Exploring Investigative Question Generation

Tom Parkhouse

School of Psychology

University of Sussex

Thesis submitted for the degree of Doctor of Philosophy in Psychology

October 2019

Declaration

The thesis is presented in 'article format' with tables and figures embedded in the text. The five empirical chapters consist of discrete articles written in a style that is appropriate for the journal they were intended to be submitted to. The opening and final chapters present overviews and discussions of the current literature and the research undertaken within the thesis.

The research submitted for this studentship was supported by the Engineering and Physical Sciences Research Council (grant number: EP/M508172/1). The author contributions for the empirical chapters are as follows: Tom Parkhouse and Thomas Ormerod were equally responsible for the conception of the research ideas and study designs. Tom Parkhouse was responsible for data collection, data analysis and the writing of the manuscripts. Tom Ormerod was responsible for providing feedback on study design and data analysis and for editing the manuscripts. The empirical chapters in this thesis have been written in a joint-authored publication format. As such, the term 'we' is commonly used. However, the writing of this thesis was entirely the authors own work.

Chapter 3 has been published in PLoS ONE: Parkhouse, T., & Ormerod, T. C. (2018). Unanticipated questions can yield unanticipated outcomes in investigative interviews. *PLoS ONE, 13*(12): e0208751. This chapter was originally written using the Vancouver referencing style, in line with PLoS ONE guidelines. For consistency, this chapter has been converted to APA referencing style for inclusion in the thesis.

I hereby declare that this thesis has not been, and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature……. Date……

Acknowledgements

Firstly, I would like to thank my wonderful supervisor, Tom Ormerod, for all of your input throughout this project. I would never have made it to the end without your positivity and support, and your ability to make sense of the complex research situations I often found myself in.

I would also like to thank the other researchers involved in the wider project: David Weir and Colin Ashby, for helping conceptualise the wider-reaching goals and implications of the project. Also, Graham Hole, who provided me with yearly feedback and advice, and always reassured me that I was on the right track. Additionally, I would like to thank Ryan Scott, who was my supervisor at Undergrad and Master's level, for helping me to realise that I wanted to go into research.

I would like to thank the admin and technical support team in the Sussex psychology department, especially Pennie, Dan, and Mar, for the countless times they helped me (usually at the last minute!) to sort out the resources and equipment that I needed.

I am hugely indebted to my fellow PhD students and friends who took part in my experiments, either as interviewees or interviewers: Anne, Geoff, Jenny, Kate, Hev, Sanj, Heather, Alex, Yasin, Chris, Charlotte, Tommy, James, and Abi- thank you so much for giving up your time to help me out.

Thank you to my family, who have always been supportive and encouraging ever since I made the decision to leave my job, get some A-Levels and go to uni. Also, without your financial support (which I know none of you could really afford!) I would never have made it this far.

I would also like to acknowledge my all of the friends who have distracted me, bought me drinks and made me laugh throughout the past four years.

Finally, I would like to thank my girlfriend, Emily, who has put up with me and kept me going throughout this project. In what has been a very stressful last year, you have always been the first to provide encouragement and support and have pushed me to the end. You're the best!

UNIVERSITY OF SUSSEX
TOM PARKHOUSE
PhD PSYCHOLOGY

EXPLORING INVESTIGATIVE QUESTION GENERATION

SUMMARY

A failure to conduct effective investigative interviews can have drastic consequences such as wrongful convictions or the inability to prevent terrorist attacks. One method of judging the efficacy of an interview is the ability to distinguish between honest and deceptive interviewees. Many techniques claim to improve the ability to detect deception, such as the CCE technique. However, little research has focused on the conditions that might enhance the ability to generate investigatively useful questions.

In series of experiments we sought to identify the underlying dimensions of question quality. Initially, we investigated unexpectedness, finding that it was a useful dimension but dependent on the content of the questions (Chapters 2 and 3). In Chapter 5 we used a bottom-up approach to develop a 3-dimensional model of question quality. These dimensions were investigative relevance, unpredictability and type of knowledge probed. The model proved to be effective in predicting the outcome of real-world investigate interviews (Chapter 6).

We also aimed to investigate the factors which might affect question generation ability. The scope of episodic information inherently available to the interviewer was shown to be a context-dependent factor affecting the ability to generate useful questions (Chapters 4 and 5). Training, via a short instructional video, was also shown to improve question generating ability (Chapter 4). Additionally, the veracity of the interviewee and the expertise of the question generator affected ability, though this was only detected by novice judges (Chapter 5).

The findings presented in this thesis have implications for the investigative community, suggesting that deliberate attention should be paid to the phrasing of key interview questions in order to ensure that they are relevant, unpredictable and probing episodic knowledge. Additionally, the findings may inform current research that is focused on developing technology designed to assist investigative interviewers.

Table of Contents

List of Appendices

Chapter 1: General Introduction

Investigative interviewing takes many forms, including police investigations, aviation security screening, benefit and insurance fraud investigations, fire scene investigations, and immigration enforcement. In each instance, the ability to ask the right questions, and generate useful information, is of utmost importance. The introduction of the PEACE model for investigative interviewing laid the groundwork for a host of techniques and methods designed to assist interviewers and improve their ability to conduct good-quality investigations. The model highlighted several crucial insights, such as the importance of building a rapport with the interviewee, thoroughly planning and preparing for interviews and, perhaps most importantly, distinguishing between different types of questions and indicating when it is appropriate, or inappropriate, to use them. This was a positive step forward, and research suggests it led to improvements in interview outcomes (Griffiths & Milne, 2006; Mcgurk, Carr, & Mcgurk, 1993). However, less is known about how to design the content of questions that are likely to yield the most useful information.

One way in which the outcomes of investigative interviews have been measured is in the ability to discriminate between truthful and deceitful interviewees. Research suggests that both the general public and trained investigators rarely demonstrate veracity detection rates greater than chance (Bond & DePaulo, 2006). There are a number of proposed reasons for this, but arguably the most persuasive is the idea that people place too much faith in non-verbal indicators, which have repeatedly been shown to be unreliable cues to deceit (Sporer & Schwandt, 2007). As such, there has been a move in recent years towards methods of interviewing that focus on eliciting verbal cues to deception. Approaches such as the Unanticipated Questions technique (UQ; Vrij et al., 2009), the Strategic Use of Evidence technique (SUE; Granhag, Strömwall, &

Hartwig, 2007), and the Controlled Cognitive Engagement technique (CCE; Ormerod & Dando, 2015) all claim to be reliable methods for distinguishing between the verbal behaviour of truth-tellers and deceivers. Each has empirical support, showing that successful interview outcomes can be reliably obtained when the techniques are performed correctly. However, to date, no published research has investigated the factors that contribute towards an individual's ability to implement these techniques and generate the types of questions they promote. Therefore, the present thesis sought to explore the potential factors which enhance or inhibit an individual's ability to generate investigatively useful interview questions. The three main research questions were:

1) *What are the dimensions by which good- and poor-quality questions differ?*
2) *Is it possible to rate those dimensions objectively?*
3) *What factors enhance or inhibit one's ability to generate good-quality questions?*

In this thesis, a series of studies, presented as a series of standalone papers, explore putative answers to these three questions.

Investigative Interviewing

An investigative interview can be defined as any interview in which an interviewer is attempting to elicit evidence or information from an interviewee, in order to assist an investigation into an incident or set of incidents or events. This is most commonly associated with police officers, who conduct investigative interviews with suspects, victims and witnesses of potential criminal behaviour. However, there are a variety of other situations in which investigative interviews are conducted, such as in benefit or insurance fraud investigations, whereby an investigator may interview a claimant to establish whether their claim is valid; after a fire, whereby a fire

investigation officer may interview the occupants of the building in order to establish whether the fire was started deliberately; or as part of routine aviation security screening in order to ensure that any passengers who are attempting to conceal their true identity or intent are apprehended.

Generally, in the UK, the purpose of an investigative interview is to gather as much useful information from an interviewee as possible. This contrasts with the US, in which police suspect interviews often take the form of interrogations designed to elicit a confession from the suspect (Gudjonsson & Pearse, 2011). Police officers in the US, and other countries, are routinely trained in the Reid Technique of interrogation (Inbau, Reid, Buckley, & Jayne, 2011). The technique incorporates nine steps which aim to provide the interrogator with a range of investigative tools, such as how to confront the interviewee, how to handle denials and objections and what to do if the suspect is passive. The technique distinguishes between interviews, which are used to establish the likelihood that the suspect is guilty, and interrogations, which are subsequently used to elicit a confession (Memon, Vrij, & Bull, 2003). Police interviewers are taught to look out for indicators of guilt during the interview, and, if such indicators are identified, the interview becomes accusatory in style, with the goal of obtaining a confession (Cleary & Warner, 2016).

Some of the methods incorporated in the Reid Technique have been questioned, with some arguing that it increases the risk of suspects making false confessions due to coercive practices (Gudjonsson, 2006; Kassin, 2006; Memon et al., 2003). Furthermore, the technique relies on interviewers determining an interviewee's veracity based on five non-verbal cues that regularly fail to find empirical support (Blair & Kooi, 2004; DePaulo et al., 2003; Vrij, 2004). In the UK, an investigative interviewer's role is not to elicit a confession, nor to determine guilt or innocence, but is instead to gather useful

information from an interviewee with which to pursue a thorough investigation. An investigative interview can be considered a success if either an uncoerced confession has been willingly given or if the interviewer is satisfied that the interview has been conducted as thoroughly and ethically as possible, leaving no information open to conjecture (Walsh & Bull, 2010).

Achieving a successful interview outcome is essential in forensic contexts, as failure to do so can have crucial implications. For example, in 2004, Sam Hallam, an 18-year-old from London, was wrongfully jailed for murder. There was no forensic evidence linking him to the crime; the entire case rested on the statements of two witnesses, which were later shown to be unreliable and inconsistent. Sam Hallam had his conviction overturned and was released in 2012, by which point he had served seven years in prison for a crime he did not commit (Laville, 2012; Poyser & Milne, 2015). Miscarriages of justice such as these might be preventable if investigative interviews are conducted correctly and thoroughly. As well as avoiding wrongful convictions, the importance and implications of obtaining a positive interview outcome are wide reaching, such as obtaining sufficient evidence to convict a guilty suspect, preventing those with criminal intent from crossing borders, or gathering useful, accurate information in order to swiftly deal with potential terrorist threats (King & Dunn, 2010).

In attempt to improve the standard of investigative interviewing, the UK police service established the Police and Criminal Evidence Act of 1984 (PACE) which introduced a range of guidelines, including some for interviewing, designed to instil an ethical approach to investigation. In 1993 the PEACE model for interviewing was implemented. PEACE ushered in a move away from interrogative-style interviews designed to elicit a confession, towards the current investigative approach designed to gather useful information (Clarke, Milne, & Bull, 2011). This followed Baldwin's

(1993) enquiry into police interviewing which concluded that over a third of interviews were not conducted satisfactorily. Amongst the issues identified by the enquiry was a failure to establish relevant facts, a lack of planning and poor technique. Research suggests that adherence to the PEACE model fosters a higher standard of ethical practice, and can improve the quality of the interview, when conducted correctly (Clarke & Milne, 2001; Griffiths & Milne, 2006; Mcgurk et al., 1993; Walsh & Milne, 2008).

One of the key differences between the interrogative interview styles, such as the Reid Technique, and the information gathering interview styles, such as PEACE, is the reliance on detecting deception. The PEACE model does not train to detect deception, unlike the Reid Technique. One major advantage of this is that it better upholds the presumption of innocence of interview suspects. The Reid Technique teaches interviewers to attempt to determine the veracity of the interviewee's account and, if they judge it to be deceptive, the interview should at that point become accusatory (Cleary & Warner, 2016). As such, the presumption of innocence is clearly lost at this stage. Kassin, Goldstein, and Savitsky (2003) showed that when interviewers are primed to believe that a suspect is more likely to be guilty, it led them to pursue more interrogative tactics and push harder for a confession, compared to interviewers who were primed to believe that the suspects were more likely to be innocent. Moreover, it also led independent observers of the audio recordings of those interviews to state that the interviewees were more defensive and, therefore, more likely to be guilty. This shows the potential harm that come from losing the presumption of innocence. However, the ability to elicit cues to deception in an investigative interview still remains important in various forensic contexts and can be used to assist a thorough investigation.

The PEACE model incorporates five key stages: Planning, Engage and explain, Account, Closure, and Evaluation. The Planning stage suggests that interviewers should create a written plan that includes the range of relevant topics to be covered, the points that must be addressed in order to prove the potential offence, and a review of any material or information that may assist the investigation. The Engage and explain phase is designed to encourage the interviewee to engage in conversation. This includes the interviewer clearly stating the reason for the interview and outlining the objectives. One intention of this stage is to build rapport with the interviewee. During the Account stage, the interviewer is attempting to gather the interviewee's account of the incident in question, usually through the use of open-ended questions. Interviewers are encouraged not to interrupt the interviewee at this stage. However, they are encouraged to clarify and expand on the interviewee's account, with more specific-closed questions, in order to fully cover each of the relevant topics introduced. In the Closure stage of the interview, the interviewer is required to summarise the interviewee's account and confirm that there is nothing to add or clarify. Finally, in the Evaluation phase, the interviewer should review the statement in terms of the wider investigation and determine the best course of action.

One of the key insights brought into focus by the introduction of the PEACE model was the distinction made between types of question. The model distinguishes between open-ended, specific-closed, forced-choice, multiple, and leading questions, and indicates when each are, and are not, appropriate. A line of questioning can be considered appropriate if it encourages the interviewee to provide a full, accurate statement, or probes for more specific, useful details within the interviewee's statement. In contrast, questions may be considered inappropriate if they limit the interviewee's ability to provide a complete account in some manner (Griffiths & Milne, 2006). For

example, the use of open-ended questions is suggested at the start of interviews in order to allow the interviewee to provide a full, open account of the incident in question. The use of specific-closed questions is encouraged when following up on information provided during the open account. The use of leading questions is warned against and only suggested as a last resort, given that they can negatively influence the interviewee's response or distort their memory (Loftus, 1975). Griffiths & Milne (2006) argue that there are 'productive' questions, such as the open-ended and specific-closed types, and there are 'unproductive' questions, such as the leading and forced-choice types. Oxburgh, Myklebust, and Grant's (2010) review of question use supports the assertion that these productive question types are more appropriate and beneficial to investigations.

The PEACE model provided the first step towards a more ethical and thorough approach to investigative interviewing. This was a hugely important step given the issues that the police were facing regarding the efficacy of interviewing prior to its induction (Baldwin, 1993; Poyser & Milne, 2015). However, the PEACE model was designed to provide a broad framework for interviewing. It does not provide specific information regarding the content of the questions used in interviews, nor does it provide details regarding the ability to detect deception. Determining an interviewee's guilt or innocence is not required by a police interviewer. However, it is important, in terms of the ongoing investigation and potential court proceedings, that the information gathered in interviews is useful to an independent observer who may be responsible for making a decision regarding the veracity of the interviewee's account. Moreover, there are many contexts other than police interviews, such as security screening, in which determining the veracity of an interviewee's account in real time is fundamental to the success of the interview. As such, a large body of research in the years since the

introduction of PEACE has focused on investigating specific interview techniques designed for this purpose.

Detecting Deception

The need for investigative interview methods that are capable of distinguishing between true and false accounts is of crucial importance in forensic investigations. It can be the difference between an innocent individual being acquitted or convicted; a guilty individual being charged or released; as well as helping to prevent potential crimes, ranging from minor infractions to large-scale attacks (King & Dunn, 2010). As such, it is vital that techniques used during police interviews, as well as other investigate interviews, are able to effectively elicit differences between honest and deceptive accounts. In terms of UK police investigations, it is independent observers, such as jury members and judges, who review the information obtained during an interview in order to inform their decision regarding the veracity of the interviewee's statement. However, other situations require the interviewer to be able to accurately determine veracity during the course of the interview. For example, in aviation security screening, the interviewer only has a brief opportunity to question each passenger, during which time they must make a decision regarding the authenticity of the passenger's identity and intent. An inability to detect deceit in this situation can have disastrous consequences, as highlighted by the September 11th 2001 terrorist attack.

Whilst such importance is placed on the ability to detect deception, research suggests that humans tend to perform poorly. Bond and DePaulo's (2006) meta-analysis of deception research revealed that, across over 200 studies, the overall veracity detection accuracy rate was just 54%. Whilst this was statistically greater than chance, it suggests that individuals are worryingly poor at detecting deceit. Moreover, there was

no difference in overall accuracy rates found between studies which employed novices and those which employed trained professionals. This was supported by Aamondt and Custer's (2006) meta-analysis showing there to be no significant difference in accuracy between students (54%) and police officers (55%), suggesting that those responsible for conducting investigative interviews are equally poor at detecting deception.

One reason for this apparent inability is the fact that individuals seem to put their faith erroneously in the efficacy of non-verbal indicators. The Global Deception Research Team's (2006) cross-cultural investigation into people's beliefs surrounding cues to deception showed that 'gaze aversion' was the most commonly cited cue, being included in almost two thirds of responses. Yet, studies (e.g., Glenberg, Schroeder, & Robertson, 1998; Wiseman et al., 2012) have shown that gaze aversion aids recollection by interviewees, and there is little evidence to suggest that eye movements are a reliable cue to deceit. Additionally, over a quarter of responses included mention of 'nervousness'. Most investigative interview contexts are situations that are likely to create some amount of anxiety. Whether being interviewed as a suspect of a crime, a witness of a crime, or simply passing through aviation security screening, each of these scenarios involves a certain level of scrutiny that is naturally going to elicit a nervous response. Therefore, conflating nervousness with deception is problematic. Moreover, these beliefs are often also held by trained professionals. Strömwall and Granhag's (2003) investigation into police officers' beliefs showed that around 60% of their sample believed that liars maintain less eye-contact and make more body movements than truth-tellers.

Whilst beliefs about the efficacy of non-verbal cues are widespread, there is little empirical support for their utility. A number of meta-analyses looking into the behaviours exhibited by truth-tellers and liars have shown there to be no evidence that

liars are more likely to avert their gaze or make more body movements than truth-tellers (DePaulo et al., 2003; Sporer & Schwandt, 2007). Due to these findings, there has been a move in recent years towards examining verbal cues to deception instead. DePaulo and colleagues' (2003) meta-analysis revealed that liars' statements tend to be briefer, less consistent and less coherent than truth-tellers' statements, suggesting that verbal cues to deceit may be more discriminating than non-verbal cues. Furthermore, several studies have shown that liars tend to respond with less detail than truth-tellers (Sooniste, Granhag, Knieps, & Vrij, 2013; Vrij et al., 2009).

The reason why verbal cues to deceit are more reliable is potentially due to the different tactics employed by truth-tellers and liars in interviews. Hartwig, Granhag, and Strömwall (2007) argue that when a deceptive interviewee is under pressure, they often use the tactic of keeping their statement simple and low in detail. This, in theory, reduces the chance that they might be challenged on any details they provide and reduces the risk that any evidence subsequently revealed by the interviewer will contradict details provided previously in their statement. In contrast, honest interviewees tend to believe that their innocence will be clear to the interviewer (Gilovich, Savitsky, & Medvec, 1998), and tend to use the tactic of providing an open, honest account rich with detail (Strömwall, Hartwig, & Granhag, 2006). Additionally, these verbal differences may be made more apparent due to the increased cognitive load experienced by deceptive interviewees. Lying is known to require greater mental effort than telling the truth (Ströfer, Ufkes, Noordzij, & Giebels, 2016). One reason for this is that an individual who is lying is required to simultaneously maintain two mental representations of an incident; they must inhibit the sequence of events that actually took place, whilst presenting the fabricated sequence of events that did not take place.

This is more cognitively demanding than telling the truth (Debey, Ridderinkhof, De Houwer, De Schryver, & Verschuere, 2015).

Research shows that, when an effective interview method is employed, it is possible to distinguish reliably between true and false accounts on the basis of verbal behaviour. Hartwig and colleagues' (2011) experiment looking into verbal behaviour differences found that the deceptive interviewees' statements were less consistent than truth-tellers', that is, their statements were more likely to contradict the known facts of the case in question. This effect was strengthened when the interviews used specific, evidence-based probes as opposed to free recall questions. Note, this notion of consistency is not the same as the type of consistency often referred to in deception research, whereby an interviewee's statements might be matched against another's. Hartwig, Granhag, and Luke's (2014) meta-analysis into the SUE interview technique showed that deceptive interviewees are more likely to make contradictory statements than truth-tellers, and this effect can be greatly emphasised if evidence regarding the incident in question is revealed late in the interview. In summary, it is now well established that verbal cues to deception are more reliable indicators than non-verbal cues. Deceptive interviewees tend to provide shorter, more contradictory, less detailed, and less consistent statements than truth-tellers (DePaulo et al., 2003) and these differences can be emphasised by certain interview techniques, such as SUE, as well as UQs and CCE (Hartwig et al., 2014; Ormerod & Dando, 2015; Vrij et al., 2009).

Whilst each of the three techniques mentioned above have similarities, there are important distinctions to be made between them. SUE (Hartwig et al., 2014) involves the strategic delaying of evidence revelation until late in an interview, which is designed to prevent the interviewee from verbal manoeuvring, by committing them to an account given before the evidence is revealed, against which responses to evidence can be

judged for inconsistencies. In CCE interviewing (Ormerod & Dando, 2015), evidence (if available) is revealed tactically throughout the interview. Like SUE, interviewees give a verbal account before any evidence is revealed, but unlike SUE, in CCE the revelation of evidence is incremental, each bit of evidence coming in response to an information gathering question. UQs (Vrij et al., 2009) is neutral on the timing of unanticipated questions. Instead, the focus for this technique is on the question content, not on the timing of evidence.

Unanticipated Questions

Another technique that has been applied to investigative interviewing is the UQ approach developed by Vrij and colleagues (2009). The technique aims to exploit the difference in cognitive load faced by honest and deceptive interviewees by preventing deceptive interviewees from relying on pre-prepared responses to anticipated questions. Research suggests that another interview tactic deployed by liars is to attempt to anticipate the questions they might be asked during an interview and prepare a set of responses. This allows them to present their lies in a way that seems more spontaneous and plausible (Colwell, Hiscock-Anisman, Memon, Rachel, & Colwell, 2007; Vrij, Fisher, & Blank, 2017). However, this tactic relies on them correctly anticipating the questions that they will be asked. Therefore, if an interviewer asks questions that they have not anticipated, they are no longer able to stick to the responses that they have prepared, and instead must respond spontaneously. Vrij (2014) argues that this increases the cognitive demand on a deceptive interviewee. In contrast, an honest interviewee should be able to answer based on their actual memory of the event, regardless of whether the question is unexpected or not.

There is empirical support for the efficacy of the technique, showing that UQs can elicit differences in verbal behaviour between truth-tellers and liars. For example, studies have shown that liars provide more detailed responses to anticipated questions than truth-tellers, but less detailed responses to UQs (Lancaster, Vrij, Hope, & Waller, 2013; Shaw et al., 2013). Mac Giolla and Granhag's (2015) investigation into the UQs approach found that liars gave longer responses to anticipated questions than truth-tellers, but shorter responses to UQs. Vrij, Mann, Leal, and Fisher (2012) found that liars' responses to UQs were rated as less plausible sounding than truth-tellers'. Taken together, the findings, across a number of studies, indicate that the UQ approach is effective in eliciting a number of verbal cues to deception.

Despite these positive findings, a number of studies have found somewhat contradictory results. Vrij (2014) asserts that, for liars, answering an expected question should be markedly easier than answering an unexpected question, given that the latter requires them to spontaneously provide an answer. However, an honest interviewee should experience a similar level of difficulty regardless of whether they have anticipated the question or not. This assumption was supported by the findings of Mac Giolla and Granhag (2015), which showed that there was no difference in self-reported difficulty between truth-tellers and liars with regards to answering anticipated questions. However, there was a difference found between ratings for the unanticipated questions, with liars stating that they found them significantly more difficult to answer than truth-tellers. Nevertheless, several other studies have shown no difference between truth-tellers and liars in terms of the level of reported difficulty in answering UQs (Granhag, Mac Giolla, Sooniste, Strömwall, & Liu-Jonsson, 2016; Sooniste et al., 2013; Sooniste, Granhag, Strömwall, & Vrij, 2014, 2015). The inconsistency of these findings presents a challenge for the cognitive load theory offered by Vrij (2014).

Despite contradictory findings regarding cognitive load, the UQs approach was a positive step in the pursuit of an interview technique that is capable of distinguishing between true and false accounts on the basis of verbal behaviour. The empirical support, showing that, in response to UQs, liars tend to provide answers that are shorter, less detailed and less plausible than truth-tellers, provides a valuable contribution to the forensic field, and helped to form the basis of more recent interview techniques such as CCE.

Controlled Cognitive Engagement

The CCE method (Ormerod & Dando, 2015) is one of the more recently developed interview techniques. It offers one of the most effective methods for detecting deception currently available. In a large-scale field study of aviation security screening methods, Ormerod and Dando showed that individuals trained in CCE can achieve veracity detection accuracy rates exceeding 70%. Moreover, unlike the design of most deception studies, where there the odds of correctly identifying veracity by chance are 50/50, in Ormerod and Dando's study, deceptive passengers were mixed in a ratio of 1:1000 with genuine passengers. CCE interviews involve brief, informal conversations across a variety of topics and works on the principle of veracity testing. The method has three phases, embodying six, empirically tested techniques, as outlined below.

The first technique employed by CCE is evidence-based veracity testing. This involves the interviewer allowing an interviewee to provide a full account, before comparing their response to the known evidence and challenging them on any inconsistencies. This approach underpins the SUE interviewing technique (Granhag et al., 2007), which has been shown to be a successful method of detecting deception

(Luke et al., 2016). Secondly, the technique promotes the use of open questions, allowing the interviewee to provide a rich verbal account. This has also been shown to be contribute effectively to veracity detection (Oxburgh & Dando, 2011). The third principle is the use of tests of expected knowledge, whereby the interviewer challenges the interviewee on an aspect of their account which, if their account is true, they should be able to answer. This has been demonstrated by Blair, Levine, and Shaw (2010) to be a useful tactic in investigations. The fourth technique is to restrict a deceptive interviewee's ability to verbally manoeuvre the conversation (Taylor et al., 2013). The fifth technique involves the use of UQs, with the intention of increasing the cognitive load of deceivers, which has been shown to emphasise verbal differences between honest and deceptive interviewee's (Vrij et al., 2009). Finally, the method instructs the interviewer to focus on the verbal content of the interviewee's responses. Research shows that there are differences between truth-tellers' and liars' verbal behaviour in response to interview questions, such as differences in response length or number of unique words used (Morgan, Rabinowitz, Hilts, Weller, & Coric, 2013).

The six techniques described above are incorporated into one overarching method that has three phases: building rapport, information gathering and veracity testing. During the initial rapport building stage, the interviewer asks neutral, non-accusatory questions that any interviewee, regardless of intent, will be able to answer honestly. This stage is used, firstly, to build rapport between the interviewer and interviewee, which has been shown to be an important aspect of investigative interviewing and can lead to better outcomes (Collins, Lincoln, & Frank, 2002; Colwell, Hiscock, & Memon, 2002). Secondly, this stage allows the interviewer to establish a baseline regarding each interviewee's behaviour when they are not under challenge. The interviewee's behaviour in the baselining phase can then later be compared to their

behaviour when they are being challenged. Baselining such as this has been demonstrated to be effective in interviews (Frank, Yarbrough, & Ekman, 2006).

The information gathering phase, involves the interviewer asking open questions and encouraging the interviewee to provide a rich, uninterrupted account, before going on to ask more focused, closed questions in order to gather more specific information, similar to the style of questioning suggested by the PEACE model (Griffiths & Milne, 2006). In an aviation security scenario, whereby there is no fixed incident under investigation, this phase can focus on any given topic, such as the interviewee's employment, family, or education. The purpose is to encourage the interviewee to reveal some information about themselves, which can be subsequently challenged in the final veracity testing phase.

In the final phase, the interviewer takes an aspect of the information provided in the previous phase and generates one or more test questions- the answer to which the interviewee should know if the information they have provided is true. These questions are referred to as tests of expected knowledge, for this reason. Furthermore, based on Vrij's work on UQs (Vrij, 2014; Vrij et al., 2009), Ormerod and Dando (2015) argue that these tests of expected knowledge will be more effective if they are not anticipated by the interviewee. An important distinction is made in this phase between semantic and episodic knowledge. Semantic knowledge is what we generally equate to 'general knowledge'. It is the information we know to be true through learning or through schematic representations we have developed based on our combined experiences of certain events. In contrast, episodic knowledge refers to the information we know to be true because we have personally experienced that specific episode (Renoult, Irish, Moscovitch, & Rugg, 2019).

To put it in terms relevant to investigative interviewing, one may have some semantic knowledge regarding a bank robbery from reading news reports of them and could use this to identify likely elements from a robbery 'script' (e.g. masked men, threatening an employee with a weapon). However, if you had genuinely experienced a bank robbery, you may also episodic knowledge of that incident, that it would not possible to know without being there, such as the distinct smell of the assailant's cologne. Therefore, if the test questions used in CCE focus on the interviewee's episodic knowledge of the topic in question, this should increase the challenge for an interviewee who is being deceptive, given that they will not have that episodic knowledge stored, and will be forced to rely on semantic knowledge. Moreover, episodic memory retrieval requires more concentration and is not automatic, unlike semantic knowledge, making it a more cognitively demanding challenge (Taylor & Dando, 2018).

The interviewer is encouraged to examine the interviewee's behaviour in responses to these challenges, monitoring whether there is a noticeable change from their behaviour at the baseline phase. A behavioural change may include the interviewee noticeably moving from a calm to an agitated disposition, or conversely, from an excited to passive disposition. However, it is important to stress the importance of the **change** in behaviour, as opposed to the behaviour itself. In an aviation security context, the agent will cycle through phases two and three several times, each time varying the topic and the temporal perspective (i.e., varying whether the topic is discussed within the context of the interviewee's past, present or future).

Ormerod and Dando (2015) tested the technique in a large-scale field study, whereby a group of airline security agents were trained in the method and were required to use it over the course of 18 months. During that time, 204 mock passengers were sent

through the security screening process, half being screened by agents using the standard security method, and half by the agents using the CCE method. Detection of mock passengers was compared against screening of a similar sample of genuine passengers selected from recordings of interviews conducted during the trial and matched for age, gender and nationality. The results revealed that the CCE trained agents identified 24 times as many mock passengers, with accuracy rates exceeding 70%, despite the fact that only one mock passenger was sent through for every 1000 genuine passengers. This provided a great deal of support for the efficacy of the technique, establishing it as the current most successful method for detecting deception, given the exceptionally low base rate of mock to genuine passengers.

One potential issue is that the veracity testing phase is an inherently difficult skill for an interviewer to learn. It requires them to listen to the interviewee's account, pick out certain testable aspects, and then generate a question that it would be reasonable to expect the interviewee to know, but that is also challenging and unexpected. Furthermore, they are required to do this repeatedly across various topics and temporal perspectives for each interviewee. Whilst the results of Ormerod and Dando's (2015) study indicate that it is possible to learn this skill, it nonetheless presents a difficult challenge for the interviewer, and necessarily requires creativity and insight. What has not been established, at this stage, is the level of variation in quality found within the CCE test questions, and the impact this may have on subsequent outcomes.

Aims of the Thesis

The two techniques discussed in this introduction, UQs and CCE, both offer empirically supported methods for detecting deception and gaining positive interview

outcomes. As such, their contribution to the investigative field should not be underestimated. Both take the initial steps made by the PEACE model, in terms of the distinction between question types, and provide further insight into the effect that the content of questions can have on interview outcomes. UQ focusses on the unexpectedness of a question, exploiting the differences in cognitive load experienced by truth-tellers and liars (Vrij, 2014). CCE built on this, incorporating UQs into a wider framework of six empirically tested techniques. In terms of question content, CCE promotes the use of tests of expected knowledge. However, to date, no research has explored the factors involved in, and contributing towards, the generation of questions such as the ones suggested by these techniques.

The purpose of this thesis is to explore those potential factors. Judging the quality of interview questions is arguably instinctive; an expert interviewer should be able to use their experience to determine the right question to ask at the right time. However, is it possible to determine the individual dimensions that distinguish between a good-quality, investigatively useful question, and a poor-quality, unhelpful question? If it were possible, this would provide those who are researching, training in or conducting investigative interviews with a set of criteria by which to categorise and judge interview questions. For this to be effective, it must first be determined whether it is possible to objectively define any such dimensions and, if so, whether the dimensions would provide a reliable rating scheme for judging question quality that is capable of predicting interview outcomes. This would allow further exploration into the factors which might enhance or constrain an individual's ability to produce good-quality questions. To this end, a series of experiments exploring the potential dimensions of question quality were designed.

Beginning with one of the more well-researched potential dimensions of question quality, Chapter 2 provides a systematic review of Vrij and colleagues' (2009) UQs approach. A literature search was undertaken, with 16 experiments being subsequently analysed. The review investigated the difference in truth-tellers' and liars' verbal behaviour across a variety of outcome measures. Additionally, it explored the various types of UQs used across the studies. It was predicted that the UQ approach would reliably distinguish between honest and deceptive interviewees.

Chapter 3 further investigated the UQs approach. Based on a number of issues raised in Chapter 2, the UQs approach was empirically tested in order to determine the effect that the method has on the veracity detection accuracy of both interviewers, at the time of interview, and independent observers, subsequent to the interview. The cognitive load theory offered by Vrij (2014) was also investigated. Additionally, the effects of question type were examined, with a direct comparison made between UQs that focussed on the planning of an event and UQs that focussed on the spatial and temporal details associated with an event. It was predicted that UQs would improve the veracity detection accuracy of both interviewers and observers, with liars finding the UQs more cognitively challenging to answer than truth-tellers.

Chapter 4 was an exploratory investigation into interview question generation. Novice participants were shown information-gathering interview clips, in which interviewees discussed a variety of broad topics, and were asked to generate investigatively useful follow up questions. The study examined the effects of topic, temporal perspective and training on question generation. The generated questions were rated by experts for subjective quality. Additionally, given the creative nature of question generating, the experts rated the questions for novelty and utility (the standard components of creativity; Finke, 1990). It was predicted that participants who had

watched a 10-minute training video in the CCE technique would generate higher quality questions than those who had not watched the video. Furthermore, it was predicted that the creativity ratings would adequately capture question quality.

Chapter 5 was a two-part study, further exploring the question generation process. In an initial pilot study, a group of novice participants were shown a series of information-gathering interview clips regarding a specific prohibited event (the theft of an exam paper). They were asked to generate investigatively useful questions after each. Subsequently, a card sort was performed on the questions in order to identify the dimensions by which good- and poor-quality questions could be distinguished. In the main study, a new group of novice and expert participants were required to generate questions using the same video stimuli. The incidents that the interviews concerned were varied in terms of both the scope of the instructions, which were either specific or general, and veracity, either being completed honestly or deceptively. The questions were rated by experts and novices for subjective general quality. Additionally, the experts rated the questions according to the dimensions identified in the pilot study. It was predicted that the dimensions identified by the pilot study would provide a reliable rating system and that the scores would correlate with the general quality ratings. Furthermore, it was predicted that the scope and veracity with which the tasks had been carried out would affect the quality of the questions subsequently generated.

In Chapter 6, the rating scheme developed in Chapter 5 was applied to a set of questions taken from real-world investigative interviews. Transcripts were gathered from 40 successful and 40 unsuccessful CCE interviews, carried out during Ormerod and Dando's (2015) aviation security field study. The questions in the transcripts were rated by two experts for subjective general quality and using the 3-dimensional rating scheme developed in Chapter 5. The questions were coded for topic and temporal

perspective. It was predicted that ratings made using the 3-dimension rating model would positively correlate with the general quality ratings. Moreover, it was predicted that questions taken from successful interviews, where the mock passenger had been identified, would be rated higher on the 3-dimensional model than questions taken from unsuccessful transcripts, where the mock passenger had not been identified. Finally, an interaction between the topic of discussion and the temporal perspective of the topic was expected to have an effect on the quality ratings.

Chapter 2. The benefits of asking the unexpected in investigative interviews: A systematic review of current evidence.

Abstract

There is much evidence to suggest that, whether trained or otherwise, humans are poor lie detectors. Many studies have reported finding that people perform little better than chance when attempting to determine the veracity of an interviewee's account (Bond & DePaulo, 2006). Many attempts have been made to develop new, strategic interview techniques designed to improve the ability of investigators. In this review we examine the efficacy of one such technique- the unanticipated questions approach (Vrij et al., 2009). By systematically searching the literature, sixteen studies were identified which were suitable for inclusion in a qualitative review, eight of which were also included in a meta-analysis. The findings revealed that, in response to unexpected questions, liars gave answers that were shorter, less detailed and less plausible than truth-tellers. Whilst these findings appear to be encouraging, some of the findings are less convincing and the current literature leaves several substantial issues to be resolved.

Introduction

The ability to detect deception is an essential skill for any individual who conducts investigative interviews. From trained police interviewers to aviation security staff, being able to distinguish between the accounts of those who are telling the truth and those who are being deceptive is of vital importance in terms of conducting thorough investigations, maintaining the law, and preventing security threats. Despite this, there is a good deal of evidence to suggest that both laypersons and trained individuals alike perform little better than chance when attempting to detect deception (Bond & DePaulo, 2006). This suggests that traditional, commonly held beliefs concerning methods for detecting deception may be insufficient. In recent years, a variety of novel techniques have been proposed, often accompanied by impressive truth-lie classification rates. One such technique is the Unanticipated Questions approach (UQ, Vrij et al., 2009), whereby interviewers ask questions that the interviewee is unlikely to have expected to be asked in advance. Since the technique was first proposed, several studies have investigated the approach, yielding mixed results. The purpose of this review is to collate and inspect the findings of those studies, in an attempt to thoroughly examine the efficacy of the technique in detecting deception.

Bond and DePaulo's (2006) highly influential meta-analysis of studies investigating the ability to detect deception, in both lay persons and legal professionals, identified an overall truth-lie classification rate of 54%. Whilst this figure was statistically greater than we would expect from chance guessing, it nonetheless reflects a fairly meagre success rate. What accounts for our seemingly poor ability to distinguish between the truth and a lie? In a cross-cultural study by the Global Deception Research Team (2006), a large sample of participants, drawn from 58 countries, were asked to answer the question “How can you tell when people are lying?” The results revealed

that almost two thirds of all respondents included ‘gaze aversion’ in their answer. In fact, gaze aversion was the most common answer provided by participants in 51 of the 58 countries. Furthermore, ‘nervousness’ and ‘increased body movement’ were each included in the responses of over a quarter of all participants. Given these results, it appears there exists a cross-cultural belief that veracity can be determined from an individual’s non-verbal behaviour. Whilst this research was conducted on lay persons, there is evidence to suggest that the same beliefs are shared by trained investigators (Hartwig & Granhag, 2015). In fact, investigators are commonly trained to use non-verbal cues as a method for detecting deception (Blair & Kooi, 2004). Strömwall and Granhag (2003) collected questionnaire data from a large sample of legal professionals and found that over 60% of the police officers they surveyed indicated that liars maintain less eye contact than truth-tellers, whilst a similar percentage suggested that liars exhibit more body movement. Taken together, the evidence points to a deep-rooted belief, by both lay persons and legal professionals, that lies can be detected from behavioural cues, with gaze aversion and increased body movement seemingly the most commonly held signs of deceit.

Despite the prevalence of these beliefs, there is little evidence to suggest that such behavioural cues are reliable indicators of deception. Sporer and Schwandt (2007) conducted a meta-analysis investigating the efficacy of non-verbal indicators and found no evidence that liars avert their gaze more than truth-tellers. Some research even suggests that liars make significantly *more* eye contact with an interviewer than truth-tellers do (Jundi et al., 2013; Mann et al., 2013). Likewise, body movements, such as leg and hand movements, have been shown to have no positive relationship with deception (DePaulo et al., 2003). In fact, in contrast to commonly-held beliefs, it has

been argued that these cues actually have a negative relationship with deception, with liars making less body movement than truth-tellers (Sporer & Schwandt, 2007).

As a result of these findings, recent research has more commonly turned to verbal behaviour in search for reliable cues to deception. A meta-analysis conducted by DePaulo and colleagues (2003) revealed that liars' statements show significantly less consistency and coherence than truth-tellers', and that liars spend a shorter proportion of the time in an interaction talking. Furthermore, liars tended to provide fewer details in their responses than truth-tellers. This latter finding has become relatively common in the literature, with many recent studies reporting a decreased level of detail in the responses of liars compared to truth-tellers (e.g., Sooniste et al., 2013; Vrij et al., 2009). According to Hartwig and colleagues (2007) there are three main reasons why liars may be more inclined to provide a simple statement that is low in detail: to avoid key details being challenged; to avoid contradictions; and because fabricating a narrative is, in itself, cognitively demanding. In contrast, honest individuals are more likely to keep their statements focussed on the truth, as opposed to keeping the story simple (Strömwall et al., 2006). This may stem from a belief that their innocence will be clear for all to see; often referred to as the 'illusion of transparency' (Gilovich et al., 1998).

The technique which has arguably received the most attention is the UQ approach develop by Vrij and colleagues (2009). This approach is based on the assumption that, before an individual attempts to be deceptive, they will put some effort into planning how to tell the lie in order to avoid detection (Hartwig et al., 2007). As part of this planning, they are likely to anticipate a range of questions which they may be asked regarding the event and attempt to prepare suitable answers (Vrij et al., 2017). It is this anticipation and planning element that the unanticipated questions approach seeks to exploit. Asking questions that a deceptive individual has not anticipated

increases their cognitive load (Vrij, 2014). Instead of mentally referring to their planned responses, they must 'think on their feet' and provide a spontaneous answer which aligns with both their overall narrative and their answers to previous questions. In contrast, for those who are telling the truth, the questions may be unexpected, but they should have less difficulty providing an answer given that they can rely on their actual experience of the event in question.

Many positive findings have been reported since the unanticipated questions approach was first formulated. For example, the approach is reported to significantly distinguish between liars and truth-tellers in terms of the detail, length, consistency and plausibility recorded in participants' responses to the unanticipated questions (Sooniste et al., 2013; Vrij et al., 2009, Vrij, Mann, et al., 2012). Based on these differences some findings report correct classification rates exceeding 80% (Lancaster et al., 2013; Vrij et al., 2009). Despite this, there have been some less convincing findings (e.g., Knieps, Granhag, & Vrij, 2013; Mac Giolla & Granhag, 2015), as well as a good deal of variation in terms of the methodology used, and types of unanticipated questions asked. At the time of writing, no attempt has been made to systematically collate and review these moderately disparate studies. As a result of this, it is unclear to what extent the technique improves veracity detection in interviews and whether the type of unanticipated question asked has any effect on the interview outcome. Therefore, the present paper is intended primarily to address this gap, considering both the efficacy of the approach generally, as well as the effect of factors such as question type and the methods used to obtain outcome measures.

Method

Identification of Studies

A comprehensive literature search was conducted in December 2015, using the databases ASSIA, PsycINFO, Science Direct, PsycARTICLES, Scopus, Web of Science, and PQDT Global. In order to find articles relating to the use of unanticipated questions in the detection of deceit, the search terms used were (*decept** OR *deceit** OR *liar** OR *lying** OR *honest** OR *truth**) AND (*unanticipate** OR *unexpect**). This search yielded a total of 2638 articles. In addition to the database search, 97 articles which had cited Vrij and colleagues' (2009) original paper on the unanticipated questions approach and 71 articles which were cited in Vrij's (2014) review paper of investigative interviewing were also included. Leading researchers in the field were contacted with requests for relevant unpublished materials. This resulted in the inclusion of one 'in press' article, kindly provided by Erik Mac Giolla (article since published: Granhag et al., 2016). Finally, the reference lists of those studies which had satisfied the inclusion criteria in the initial search were examined to identify any relevant papers not identified previously. This resulted in the inclusion of a further 19 studies. After the removal of duplicates, there were a total of 1744 articles eligible for title and abstract screening.

Inclusion/Exclusion Criteria

The 1744 papers were subsequently screened according to the following inclusion criteria: (a) must be an empirical study; (b) must include a direct comparison of truths and lies; (c) participants must be subjected to some form of investigative questioning; (d) must include a direct comparison of anticipated and unanticipated questions; (e) degree to which questions were anticipated by participants must be measured; (f) participants must be over the age of 18.

Based on these criteria, 1654 papers were excluded during the title and abstract screening stage. This left 90 remaining papers eligible for full-text screening. Of these, a further 77 articles were excluded on the grounds that they did not meet the inclusion criteria (see Figure 1), leaving 13 papers to be included in the systematic review.

Data Extraction

Data were extracted from the included 13 papers (16 studies in total). The information extracted from these studies were: authors' names; year of publication; number of participants; mean age of participants; percentage of female/male participants; type of task used in deceptive condition; type of unanticipated questions asked; the relevant outcome measures and findings; and, where possible, the level of detail in the participants' responses to both the anticipated and unanticipated questions (means and standard deviations of detail, based on either actual amount of detail provided or on independent ratings of the level of detail provided).

Five studies (Clemens et al., 2013; Granhag et al., 2016; Mac Giolla & Granhag, 2015, study 2; Vrij, Leal, Mann, & Fisher, 2012, study 2; Vrij, Leal, Mann, & Granhag, 2011, study 2) did not measure the amount of detail in participant's responses, and so were not included in the meta-analysis. In addition to this, three studies reported the data relating to detail in a way which meant it was not possible to accurately calculate mean difference scores, and so these studies were also not included in the meta-analysis (Lancaster et al., 2013; Vrij et al., 2009; Warmelink, Vrij, Mann, Jundi, & Granhag, 2012). Therefore, 16 studies were included in the qualitative review, whilst eight were further analysed statistically in terms of the level of detail provided in participants' responses to both anticipated and unanticipated questions.

Statistical Analysis

Means and SD of the level of detail in participants' responses to anticipated and unanticipated questions were obtained from eight studies. Of these, five studies measured level of detail on a 1-7 scale (1 = very little amount of detail; 7 = very high amount of detail), whilst the remaining three studies reported the actual number of details provided. Standardised mean differences were calculated using Cohen's *d* statistic in order to reflect the mean difference in the level of detail provided in response to anticipated questions and unanticipated questions, divided by their shared standard deviation. Two studies (Vrij, Leal, et al., 2012, study 1; Vrij et al., 2011, study 1) used a within-participants design, whereby participants took part in both the honest and deceptive conditions at separate times. However, these were treated as between-participant studies when calculating the mean difference scores, as recommended by Dunlap, Cortina, Vaslow, and Burke (1996). A random effects model (Hedges & Olkin, 1985) was used to conduct the meta-analysis, whilst the restricted maximum likelihood method was used to estimate between study variance.

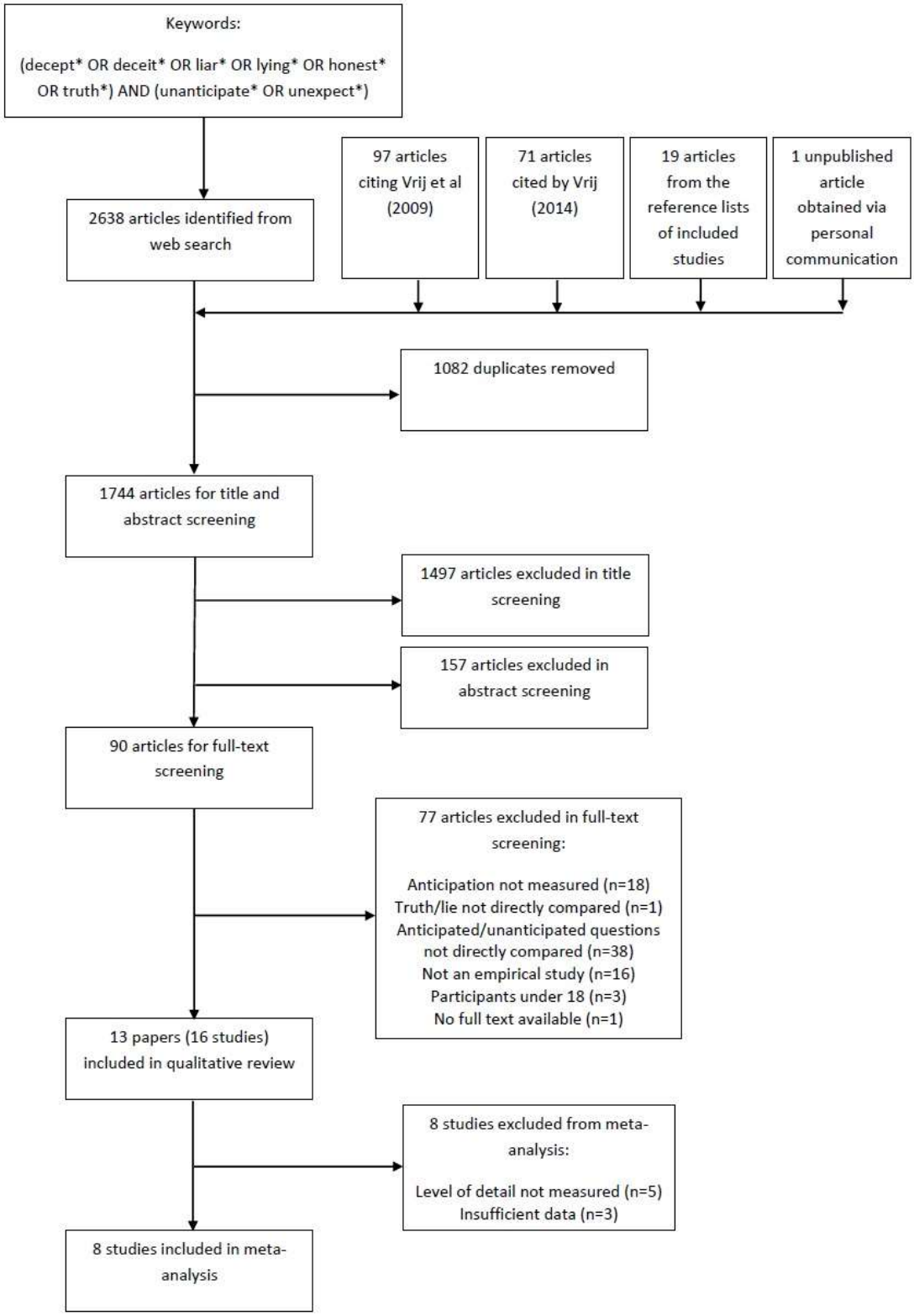

Figure 1. Flow diagram displaying the search and selection process for studies to be included in both the qualitative review and meta-analysis.

Results

Qualitative Review

Thirteen papers (totalling 16 studies) were included in the qualitative review, having met the required inclusion criteria. Across these studies there were a total of 1538 participants. The mean age of the participants overall was 27.38 years and 64% were female.

Initially, this review will assess the variation in the content of the UQs, before examining the reported outcome measures. Numerous outcome variables were measured across the 16 studies. This review will cover five of these: difficulty in answering the questions; level of detail in responses; length of responses, plausibility, and accuracy of independent observers' veracity judgements. Additionally, each of the 16 studies measured the level to which participants anticipated being asked the questions. In all of the studies included in this review, the participants indicated that the questions which were intended to be 'unanticipated' did indeed come as a significantly greater surprise than the 'anticipated' questions.

Content of questions. The first issue to address is the type of unanticipated questions used by the studies. Ten used questions that focussed on the planning of a particular task (Clemens et al., 2013; Granhag et al., 2016; Mac Giolla & Granhag, 2015, studies 1 & 2; Sooniste et al., 2013, 2014, 2015; Vrij et al., 2011, studies 1 & 2; Warmelink et al., 2012). For example, Sooniste and colleagues (2015) had their participants plan either a legal protest (truth-tellers) or an illegal protest (liars). Liars were additionally instructed to create a 'cover story' to disguise their criminal intent. Before carrying out the protest, the participants were intercepted and informed that they would be interviewed. Liars were instructed to use their cover story in the interview,

whilst the truth-tellers were simply told to tell the truth. The interview contained questions concerning the intentions of the participants (anticipated questions), as well as questions relating to their planning of the task (unanticipated questions). Examples of questions based around planning include ‘what was the main goal of your planning?’; ‘What was the final thing you planned?’; and ‘Did you have an alternative plan . . . in case things were to go wrong?’

Five of the studies used questions that focussed on the spatial or temporal details involved in the task (Lancaster et al., 2013; Shaw et al., 2013; Vrij et al., 2009; Vrij, Leal, et al., 2012, studies 1 & 2). For example, in the interviews conducted by Vrij and colleagues (2009), they asked both specific spatial questions (e.g., “In relation to the front door, where did you and your friend sit?”) and specific temporal questions (e.g., “In which order did you discuss the different topics you mentioned earlier?”) Finally, one study used a request to provide a sketch as the unanticipated question. Vrij, Mann, and colleagues (2012) had participants describe either their real place of work (truth-tellers) or a fake place of work (liars). In interviews, the participants were asked to verbally describe this workplace (anticipated question) and to provide a sketch of the workplace (unanticipated question).

Difficulty. Seven of the studies measured self-reported difficulty in answering the questions. All seven reported finding that the unanticipated questions were perceived as more difficult to answer than the anticipated questions. Three found this to be the case without a question type × veracity interaction (Sooniste et al., 2013, 2014, 2015), indicating that the unanticipated questions were more difficult to answer for liars and truth-tellers alike. However, four did report an interaction. Mac Giolla and Granhag (2015), in both study 1 and 2, reported that liars experienced the unanticipated questions as being significantly more difficult to answer than truth-tellers, whilst no such

difference was evident for the anticipated questions. Clemens and colleagues (2013) reported that liars found the unanticipated questions significantly more difficult to answer than the anticipated questions, whilst the converse effect was found for truth-tellers who found the anticipated questions significantly more difficult to answer. Finally, the liars in Granhag and colleagues' (2016) study found the anticipated questions more difficult to answer than the truth-tellers. However, there was no such difference between the groups for the unanticipated questions.

Detail. Eleven of the studies measured the amount of detail in the participants' responses to anticipated and unanticipated questions (Lancaster et al., 2013; Mac Giolla & Granhag, 2015, study 1; Shaw et al., 2013; Sooniste et al., 2013, 2014, 2015; Vrij et al., 2009, 2011, study 1; Vrij, Leal, et al., 2012, study 1; Vrij, Mann, et al., 2012; Warmelink et al., 2012). The method for measuring detail varied between the studies. Six used a subjective method, showing transcripts of the interviews to independent coders who rated them using Likert scales ranging from 1 (very low in detail) to 7 (very high in detail). The remaining five used a more objective method and had coders count every detail mentioned in the responses of the participants, to provide an overall total. Though, what constitutes a 'detail' is also, arguably, subjective.

As a result of this, it is also of interest to assess who applied these coding methods across the studies. For seven of the studies, one unspecified coder was responsible for rating/counting the level of detail in the transcripts with a second coder assessing a sample of the data (ranging between 10% and 50%). In two of the studies conducted by Sooniste and colleagues (2013; 2014), two assistants rated 100% of the transcripts in terms of the level of detail in responses. Finally, in both Vrij and colleagues' (2011) and Vrij, Leal and colleagues' (2012) studies, three raters were

trained to code the number of details in the transcripts using the Reality Monitoring visual detail criterion.

Looking at the effect of unanticipated questions on the level of detail in response, five of the studies found that liars were less detailed than truth-tellers in their responses to both the anticipated questions and the unanticipated questions (Mac Giolla & Granhag, 2015, study 1; Sooniste et al., 2014, 2015; Vrij et al., 2009; Vrij, Leal, et al., 2012, study 1). A further three found a significant veracity × question type interaction. Of these, two found that, in response to anticipated questions, liars' answers were more detailed than truth-tellers but less detailed than truth-tellers when answering the unanticipated questions (Lancaster et al., 2013; Shaw et al., 2013). The same effect was reported by Warmelink and colleagues (2012). However, in this instance, liars were only significantly less detailed than truth-tellers in response to one of the three unanticipated question types used (transportation). Similar to these findings, two of the studies reported that liars' responses to the unanticipated questions were significantly less detailed than the responses of truth-tellers, though the groups were equally detailed in response to the anticipated questions (Sooniste et al., 2013; Vrij, Mann, et al., 2012). Converse to the majority of the findings reported here, one study (Vrij et al., 2011, study 1) found that, whilst liars gave less detailed answers than truth-tellers in response to anticipated questions, the groups were equally detailed when answering the unanticipated questions.

Response Length. Three of the studies measured the overall number of words recorded in the participants' responses to the interview questions. In study 1 of Mac Giolla and Granhag's (2015) paper, they found that truth-tellers gave lengthier responses than liars to both anticipated and unanticipated questions. Whilst they reported a significant veracity x question type interaction, this was accounted for by the fact that

the difference between the groups was much more pronounced in the unanticipated questions. In study 2 of the same paper a similar effect was reported. However, in this case the interaction could be accounted for by the finding that truth-tellers used more words in response to unanticipated rather than anticipated questions, whilst the opposite was true for liars who gave longer responses to the anticipated questions. Finally, Sooniste and colleagues (2013) found no difference in length between liars' and truth-tellers' responses to anticipated questions, but found that truth-tellers gave significantly longer answers to the unanticipated questions.

Plausibility. Three studies measured the plausibility of the participants' answers (Vrij, Leal, et al., 2012, study 1; Vrij et al., 2011, study 1; Vrij, Mann, et al., 2012). Each of the three studies measured plausibility by providing transcripts of the interviews to three unspecified coders who rated the accounts on a Likert scale ranging from 1 (not at all plausible) to 7 (very plausible). The term 'plausibility' was defined as the extent to which the coder could imagine the interviewee taking the route they had discussed, or working in the place they had described, making this a somewhat subjective outcome measure in nature.

Two of the studies found a main effect of veracity, with truth-tellers' accounts being perceived as significantly more plausible than liars' accounts in response to both the anticipated and unanticipated questions (Vrij, Leal, et al., 2012, study 1; Vrij et al., 2011, study 1). Vrij, Mann, and colleagues (2012) also found a significant effect of veracity, with truth-tellers' accounts reported as being more plausible than liars'. However, when looking at the question types separately, it was shown that truth-tellers' answers were only significantly more plausible than liars' in response to the unanticipated questions, with responses to the anticipated questions being equally plausible.

Veracity Judgement Accuracy. Finally, two studies measured the accuracy of the veracity judgements made by untrained participants (Vrij, Leal, et al., 2012, study 2; Vrij et al., 2011, study 2). In these studies, a group of participants were given separate transcripts of the anticipated and unanticipated interview questions. Both studies asked the participants to rate whether they thought the interviewee was lying on a Likert scale ranging from 1 (definitely not lying) to 7 (definitely lying), whilst in Vrij and colleagues' (2011) study they were also required to provide a dichotomous truth/lie judgement about each transcript. The participants in Vrij, Leal, and colleagues' (2012) study gave significantly higher ratings to liars' statements (i.e., made more accurate veracity judgements) when using transcripts from the unanticipated questions. However, no such difference was found for the anticipated questions. This effect was also found by Vrij and colleagues (2011) when analysing the Likert scale judgements. In terms of the participants' accuracy rates when making the dichotomous judgements, the findings revealed that accuracy was significantly better than chance for both truth-tellers and liars when using transcripts of the unanticipated questions, whilst this was not the case when using transcripts of the anticipated questions.

Meta-Analysis

In order to further investigate the efficacy of the unanticipated questions approach in eliciting cues to deception, a meta-analysis was conducted using the level of detail in the participants' responses as the outcome measure. This measure was chosen as it was the most commonly used outcome variable reported by the studies included in the review. Eleven of the studies measured level of detail. Three were excluded from the meta-analysis due to a lack of sufficient data, leaving eight remaining studies to be included.

As can be seen in Table 1 and Figure 2, the findings of the meta-analysis reveal that the liars' responses were significantly less detailed than truth-tellers when answering the anticipated questions (d = -0.41, p = .01), indicating that veracity had a small-to-medium effect on the level of detail given to these questions. However, the effect was more pronounced when participants were responding to the unanticipated questions, with the results again showing that liars were significantly less detailed than truth tellers when answering these questions (d = -0.84, p < .001), representing a large effect.

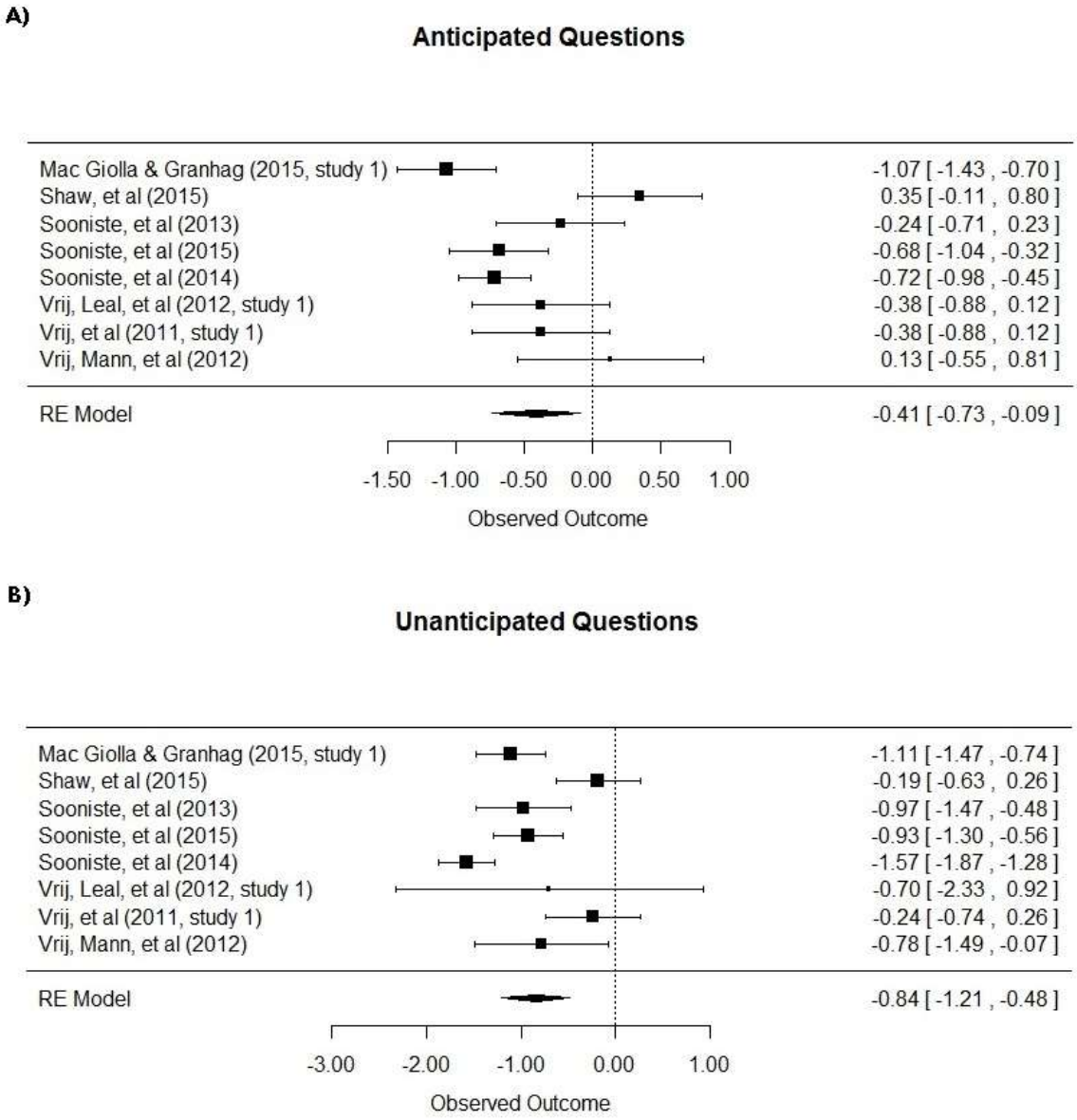

Figure 2. Forest plots showing the effect sizes in studies investigating differences in the level of detail provided by deceptive and honest participants in response to (A) anticipated questions and (B) unanticipated questions. Negative effect sizes indicate that the deceptive participants' responses were less detailed than the honest participants'.

Table 1.

Effect sizes and estimates of heterogeneity for level of detail provided in response to anticipated and unanticipated questions.

Question Type	*N*	Mean difference			Heterogeneity		
		d	*p* value	95% CI	Tau^2	I^2 (%)	Q
Anticipated	884	-0.41	.01	[-0.73; -0.09]	0.16 (0.11)	78.49	31.20***
Unanticipated	884	-0.84	< .001	[-1.21; -0.48]	0.20 (0.14)	78.76	36.94***

Note: Effect sizes are Cohen's d. A negative d indicates that liars provided a less detailed response than truth-tellers. N = number of participants; CI = confidence interval; Tau^2 = estimated amount of heterogeneity and associated standard error; I^2 = total heterogeneity/total variability; Q = test for heterogeneity; *** p < .0001

For both anticipated and unanticipated questions there was significant heterogeneity. As such, moderator analyses were warranted in order to investigate potential contributing factors. In terms of the anticipated questions, moderator analysis was conducted for the method by which level of detail was measured (overall number of details or Likert scale). This analysis was also conducted with the unanticipated questions, as well as the type of unanticipated question asked (spatial/temporal details, planning, or sketch).

As can be seen in Table 2, when assessing the effect of veracity on the level of detail given in response to anticipated questions, the method by which detail was measured was not associated with significant between-groups heterogeneity when compared to a chi-square distribution, $x^2(1, N = 2) = 2.15$, $p > .05$. However, significant within-groups heterogeneity was present, suggesting that other factors, beyond veracity

and measurement type, may exert influence on the level of detail in participants' responses.

Table 2.

Moderator analysis of impact of measurement type on the level of detail in response to anticipated questions.

Moderator	Moderator *Q* *(QM)*	No. studies	*d*	Residual *Q* *(QE)*
Detail measurement method	2.15			19.38**
Actual number of details		3	-0.13	
Level of detail (1-7 Likert scale)		5	-0.45	

Note: Effect sizes are Cohen's d. A negative effect size indicates that liars provided a less detailed response than truth-tellers. ** p <.01.

The findings of the moderator analyses conducted on the unanticipated questions can be seen in Table 3. There was significant between-groups heterogeneity for detail measurement method when compared to a chi-square distribution, $X^2(1, N = 2) = 11.76$, $p < .001$. The difference between the amount of detail provided by liars and truth tellers, in response to unanticipated questions, was larger when detail was measured on a Likert scale than when overall number of details were recorded. Finally, the type of unanticipated question asked did not significantly moderate the effect, $X^2(2, N = 3) = 2.29$, $p > .05$.

Table 3.

Moderator analyses of impact of measurement type and question type on the level of detail in response to unanticipated questions.

Moderator	Moderator *Q (QM)*	No. studies	*d*	Residual *Q (QE)*
Detail measurement method	11.76***			10.97
Actual number of details		3	-0.24	
Level of detail (1-7 Likert scale)		5	-0.89	
Unanticipated question type	2.29			22.68***
Spatial/temporal details		5	-0.29	
Planning		2	-0.70	
Sketch		1	-0.49	

Note: Effect sizes are Cohen's d. A negative effect size indicates that liars provided a less detailed response than truth-tellers. *** p <.001.

Discussion

The present review paper sought to investigate the efficacy of asking unanticipated questions in interviews as a means of eliciting cues to deception. The approach is a relatively recent addition to the field of deception research and, therefore, the empirical investigations on which to base firm judgements are limited in number. However, the outcome of this review appears to give some cause for optimism. By systematically searching the literature, 13 papers (including 16 studies) investigating the use of unanticipated questions were identified, having met the criteria required to make a series of analytic comparisons. The results showed that asking unanticipated questions results in liars providing less detailed, shorter and less plausible responses when

compared to those who are telling the truth (e.g., Lancaster et al., 2013; Sooniste et al., 2013; Vrij, Mann, et al., 2012). Taken together, these findings demonstrate the potentially promising ability of the unanticipated questions approach to provide accurate and reliable cues to deception. However, there are some issues to address.

According to Vrij (2014), asking questions that the interviewee has not anticipated in advance increases the cognitive load of those who are being dishonest, resulting in marked differences in their verbal behaviour when answering the unexpected questions compared to when answering the questions they had anticipated. Vrij further asserts that truth-tellers should not experience much difference in the level of difficulty experienced when answering expected and unexpected questions, resulting in relatively comparable verbal behaviour across the interview. If this were the case, it would be reasonable to expect that liars report finding the unanticipated questions more difficult to answer than truth-tellers. Whilst this was found in three of the studies included in this review (Clemens et al., 2013; Mac Giolla & Granhag, 2015), others found that liars and truth-tellers experienced the unanticipated questions as equally difficult to answer (e.g., Granhag et al., 2016; Sooniste et al., 2013) and in all seven studies which measured difficulty, significant main effects of question type were found. The unanticipated questions were reported as more difficult to answer than anticipated questions, for both groups. This suggests that asking unanticipated questions may in fact raise the cognitive load, not just for liars, but for truth-tellers as well. As a consequence of this, the changes in verbal behaviour expected to be exhibited by liars (Vrij, 2014) could potentially occur in truth-tellers also, presenting a significant obstacle to the efficacy of the approach.

Despite this, the review found some promising patterns across the studies regarding various potential cues to deception. In terms of the length of responses, it

seems that those who are telling the truth gave lengthier responses than liars whether answering anticipated or unanticipated questions (Mac Giolla & Granhag, 2015), suggesting that this may be a general cue and not one elicited as a direct result of question type. However, Sooniste and colleagues (2013) reported only finding differences in the length of responses to unanticipated questions, whilst in study 2 of Mac Giolla and Granhag's (2015) paper they showed that liars gave longer responses to the anticipated questions than unanticipated questions, with the converse being the case for truth-tellers. Looking at plausibility, whilst two of the studies reported that truthful accounts were more plausible than dishonest accounts, regardless of question type (Vrij, Leal, et al., 2012; Vrij et al., 2011), Vrij, Mann, and colleagues (2012) found that this was only the case when looking at the responses to unanticipated questions. Taken together, it could be suggested that asking unexpected questions might not directly produce cues to deceit but may instead work to emphasise subtle cues which are exhibited naturally by liars.

A similar pattern was found among the studies which measured the level of detail in participants' responses. Ten of the eleven studies found that liars gave less detailed answers than truth-tellers to the unanticipated questions. However, if this was also the case in response to the anticipated questions, it could not be confidently claimed that the difference between liars and truth-tellers was due to question type. This was found to be the case in half of the studies (e.g., Mac Giolla & Granhag, 2015; Sooniste et al., 2015). However, others reported no significant difference in detail between the two groups' responses to the anticipated questions (Sooniste et al., 2013; Vrij, Mann, et al., 2012), whilst some found that liars' responses to the anticipated questions were more detailed than truth-tellers' (Lancaster et al., 2013; Shaw et al., 2013).

These mixed findings were highlighted by the meta-analysis which revealed that, whilst there was a significant effect of veracity on the level of detail given to both question types, with liars' responses being less detailed in each, the effect was much larger when looking at the responses to unanticipated questions. Furthermore, the moderator analysis revealed that the effect was potentially being moderated by the method used to measure the level of detail, with a larger difference between honest and deceptive responses being observed when detail was measured using a Likert scale. Taken together, the findings suggest that the unanticipated questions approach may be an effective tool for accentuating naturally occurring cues to deception, though more robust, objective methods may be required in order to fully investigated the efficacy of the technique. However, it should be noted that the meta-analysis contained only eight studies, which is a very small sample for investigating moderator effects, and so these findings should be interpreted with caution.

Despite some of the more positive findings regarding the efficacy of the technique, it is worth considering the method used to analyse certain outcome measures. Both 'level of detail' and 'plausibility' were used to show the effects of asking unanticipated questions, with some success. But several studies measured these outcomes by asking one or two unspecified coders to rate transcripts using a Likert scale. Furthermore, there was a lack of clear definition provided for these terms across several of the studies. This raises a question about the validity of such measurements, given how subjective they inherently are. Moreover, the moderator analysis conducted as part of the meta-analysis showed that there was a larger difference in level of detail between responses from honest and deceptive participants when level of detail was measured on a Likert scale, compared to when the number of details were counted. Two of the studies (Vrij, et al., 2011; Vrij, Leal, et al., 2012) measured level of detail by

training three raters to code the transcripts according to the Reality Monitoring framework. This is, arguably, a more objective and clearly operationalised method by which to measure level of detail and should perhaps be considered the preferred method in future studies.

One issue revealed by this review is the disparity between the methods with which the approach has been investigated. There have been a whole range of distinct tasks and scenarios used by the studies included in this review, with some requiring the deceptive participants to actually carry out a task and then use a cover-story in the subsequent interview (e.g., Shaw et al., 2013; Vrij et al., 2009), some had participants plan a task but not carry it out (e.g., Clemens et al., 2013; Sooniste et al., 2013), whilst others simply asked the liars to create a false story about their occupation or travel plans (Vrij, Mann, et al., 2012; Warmelink et al., 2012). It has been argued that such differences in task type can affect the ability to detect deception, due to the differential levels of cognitive effort required to carry them out (Memon, Ormerod, & Dando, 2013). Therefore, future studies may wish to investigate whether the type of task used moderates any of the effects currently attributed to the use of unanticipated questions. The literature is also somewhat vague when it comes to the nature of the unanticipated questions themselves. As can be seen in this review, a variety of question types have been employed, including planning-based questions (e.g., Mac Giolla & Granhag, 2015; Sooniste et al., 2015), specific spatial detail questions (Vrij et al., 2009), specific temporal detail questions (Shaw et al., 2013) and even non-verbal, sketch-based questions (Vrij, Mann, et al., 2012). In addition to this, the term ‘unanticipated question’ has, to date, not been operationalised in the literature, with current definitions somewhat ambiguous. Future studies should seek to address this by attempting to provide a clear definition as to what constitutes an unanticipated question. Additionally, it may be

beneficial to conduct a direct comparison of the various types of unanticipated questions used by the studies in this review.

The key issue brought to light by this review is that not one of the 16 studies included required the interviewers to make real-time judgements concerning the veracity of the participants. Instead, the researchers focus on the various potential cues said to be attributable to the use of unanticipated questions. Whilst this has revealed some interesting, and potentially useful, findings, there is currently no indication as to whether the approach is effective in a practical application. Two of the papers included subsequent observer studies (Vrij, Leal, et al., 2012; Vrij et al., 2011), whereby a group of participants attempted to categorise transcripts of the interviews as true or false, whilst others report successful classification rates based on discriminant analyses of the level of detail in responses (Lancaster et al., 2013; Vrij et al., 2009). However, such analyses do not accurately reflect investigative settings in which the veracity of the interviewee must be determined during the interview itself. This, therefore, represents an important gap in the literature which needs to be addressed in future research.

Overall, the findings of this review are somewhat mixed. There is certainly some cause for optimism, with the majority of studies reporting that, when asked unexpected questions, liars provide shorter, less detailed and less plausible answers than those who are telling the truth. However, it is necessary to exercise a certain degree of caution. Several studies report that liars' responses to anticipated questions are also shorter and less detailed than truth-tellers', suggesting that the use of unanticipated questions may simply emphasise naturally occurring cues to deceit. Whether these cues are accentuated to a degree that can be useful in practical settings is yet to be determined. Future research should focus on establishing whether the use of unanticipated questions results in greater accuracy of real-time veracity judgements.

Chapter 3: Unanticipated questions can yield unanticipated outcomes in investigative interviews.

Paper published in PLoS ONE (2018)

Abstract

Asking unanticipated questions in investigative interviews can elicit differences in the verbal behaviour of truth-tellers and liars: When faced with unanticipated questions, liars give less detailed and consistent responses than truth-tellers. Do such differences in verbal behaviour lead to an improvement in the accuracy of interviewers' veracity judgements? Two empirical studies evaluated the efficacy of the unanticipated questions technique. Experiment 1 compared two types of unanticipated questions (questions regarding the planning of a task and questions regarding the specific spatial and temporal details associated with the task), assessing the veracity judgements of interviewers and verbal content of interviewees' responses. Experiment 2 assessed veracity judgements of independent observers. Overall, the results provide little support for the technique. For interviewers, unanticipated questions failed to improve veracity judgement accuracy above chance. Reality monitoring analysis revealed qualitatively distinct information in the responses to the two unanticipated question types, though little distinction between the responses of truth-tellers and liars. Accuracy for observers was greater when judging transcripts of unanticipated questions, and this effect was stronger for spatial and temporal questions than planning questions. The benefits of unanticipated questioning appear limited to post-interview situations. Furthermore, the type of unanticipated question affects both the type of information gathered and the ability to detect deceit.

Introduction

Bond and DePaulo's (2006) influential meta-analysis of deception detection reached a worrying conclusion: individuals, regardless of training or experience, are generally poor at distinguishing between truth and lies. Analysing the accuracy of veracity judgements made across 206 studies involving over 20,000 judges, the authors found an overall accuracy rate of just 54%, in part because the general public and trained experts alike appear erroneously to put their faith in non-verbal indicators of deception (Colwell, Miller, Lyons, & Miller, 2006; Global Deception Research Team, 2006). DePaulo and colleagues' (2003) meta-analysis revealed that statements made by liars were less consistent, less coherent, and contained fewer details than those given by truth-tellers. Thus, recent research has focussed on verbal behaviours such as differences in response length, level of detail and consistency, as cues to deceit (Hartwig et al., 2011; Sooniste et al., 2015; Vrij et al., 2009). The unanticipated questions technique (UQ; Vrij et al., 2009), evaluated in this paper, is designed to emphasise differences in verbal behaviours of truth-tellers and liars.

Asking questions that an interviewee has not anticipated should, according to Vrij (2014), increase a liar's cognitive load, resulting in observable differences in their verbal behaviours compared to those of an honest interviewee. Research has shown that liars give less detailed, less plausible, and/or less consistent answers in response to unanticipated questions than truth-tellers (Sooniste et al., 2015; Vrij, Mann, et al., 2012). Some interesting new work has shown that unanticipated questions may even be successfully utilised to detect false information being provided electronically, showing that liars exhibit certain cues to deceit, such as prolonged mouse trajectory, when answering unanticipated questions (Monaro, Gamberini, & Sartori, 2017). Numerous studies have investigated the efficacy of the unanticipated questions approach in

distinguishing between truth-tellers and liars (Lancaster et al., 2013; Warmelink et al., 2012). A recent meta-analysis showed that the cognitive approach to lie detection, which makes use of unanticipated questions, led to an overall detection accuracy rate of 71% (Vrij et al., 2017). Furthermore, unanticipated questioning is one of the six principles of the Controlled Cognitive Engagement interview technique, arguably one of the most successful practical methods for detecting deception developed for field use (Ormerod & Dando, 2015) and has been recommended as best practice in intelligence interviewing (CREST, 2016).

The majority of studies investigating the efficacy of the approach have focussed on post-hoc analyses of interviewees' verbal behaviour (e.g., statement consistency, level of detail, etc.), as opposed to real-time veracity judgements made by interviewers. Research to date has not addressed whether effects of asking unanticipated questions are noticeable to the interviewer. The present study was designed to evaluate the unanticipated questions technique, specifically whether its efficacy extends to real-time veracity judgements.

The UQ approach, it is suggested, exploits differences in the cognitive load faced by truth-tellers and liars. It is well established that telling a lie imposes greater cognitive load on the individual than telling the truth (Ströfer et al., 2016; Vrij, 2014). Results from fMRI studies have shown that lying, compared to truth telling, is associated increased neural activity in the prefrontal cortex, an area often linked to cognitive engagement (Christ, Van Essen, Watson, Brubaker, & McDermott, 2009; Kaylor-Hughes et al., 2011). There are a number of reasons why lying may be more cognitively demanding than telling the truth. For example, in an interview, a liar must present their false account while simultaneously inhibiting the truth (Debey et al., 2015). Additionally, liars are more likely to monitor and control their own outward

behaviour, while also attempting to decipher the behaviour of the interviewer, which again increases their cognitive load (Vrij, Fisher, Mann, & Leal, 2006). Interview techniques that increase the cognitive load faced by liars have been shown to improve veracity judgement accuracy rates (Dando, Bull, Ormerod, & Sandham, 2015).

Given the opportunity, liars plan how they will behave and what they will say when interviewed (Hartwig et al., 2007). As part of this planning, they are likely to anticipate questions they may be asked and prepare suitable responses to them, developing a 'lie script' (Colwell et al., 2007; Vrij et al., 2017). However, planning will only help if they correctly anticipate the questions that are asked. By asking unanticipated questions, the interviewer breaks the lie script and forces liars to answer spontaneously, which should increase their cognitive load and change their verbal behaviour (Vrij, 2014). In contrast, an interviewee telling the truth should have less difficulty providing a response to unanticipated questions because they can rely on real memories of events. Accordingly, Vrij (2014) states that 'truth-tellers experience similar levels of cognitive load while answering expected and unexpected questions, and they should produce more comparable answers to the expected and unexpected questions than liars.' (p. 187).

Clemens and colleagues (2013) argue that when liars are formulating their lie script, they tend to prepare for questions that focus on their intentions (e.g., "What items did you intend to purchase whilst at the shopping centre today?") and fail to prepare for questions about the planning of these intentions (e.g., "Tell me about the order in which you planned to purchase these items"). Sooniste and colleagues (2013) had participants plan either a non-criminal (truth-tellers) or a mock-criminal (liars) act. Liars were further instructed to prepare a cover story to mask their criminal intentions. In subsequent interviews, both groups of participants were asked questions concerning

their intentions and the planning of their intentions. The planning questions were rated as significantly less anticipated than the intentions questions. Furthermore, truth-tellers' responses to the planning questions were shown to include significantly more detail than liars' responses, with no such difference occurring in response to questions on their intentions. This supports the idea that unanticipated questions give rise to noticeable differences in the verbal behaviour of truth-tellers and liars, and subsequent studies have reported similar differences (Sooniste et al., 2015; Warmelink et al., 2012).

Other studies have focussed on spatial and temporal details as the basis for unanticipated questions (Shaw et al., 2013; Vrij, Leal, et al., 2012). Vrij and colleagues (2009) asked pairs of participants to either tell the truth or lie about having lunch together. They asked general questions about the task that might be anticipated (e.g. "Can you tell me in as much detail as possible what you did while you were in the restaurant?"), as well as specific spatial and temporal detail questions (e.g. "In relation to the front door, where did you and your friend sit?"; "How long was it between the staff taking your order and you receiving your food?"). Participants rated spatial and temporal questions as less anticipated than the general questions. Moreover, statements provided by lying pairs were less consistent than honest pairs, but only when answering the spatial and temporal questions. Liars' responses contained less detail than truth tellers' responses across all question types, and this difference was more pronounced in the spatial and temporal questions. This type of questioning has subsequently been applied to individual interviewees with similar findings (Lancaster et al., 2013).

Although the unanticipated questions approach has received considerable support in terms of its ability to distinguish true and false accounts on the basis of verbal cues (Lancaster et al., 2013; Vrij et al., 2009), there have been some inconsistent findings. Vrij and colleagues (2011) found that, while liars gave less detailed answers to

anticipated questions, there was no difference between truth-tellers and liars in the amount of detail provided in response to unanticipated questions. One potential reason for these mixed findings is variability in the types of unanticipated question used across studies. Knieps and colleagues (2013) asked interviewees unexpected questions about the occurrence of a mental image they may have had during their planning of a mock criminal event; Vrij, Mann, and colleagues (2012) study required interviewees to provide a sketch of their workplace; while Warmelink and colleagues (2012) interviews included unanticipated questions about transportation. Furthermore, Warmelink and colleagues introduced the idea of familiar and unfamiliar lies, with unanticipated questions regarding the background and details associated with interviewees' occupations. In general, studies have focussed either on questions regarding the planning of an event or on spatial and temporal details associated with an event. Although it is reasonable to imagine that different forms of unanticipated question will elicit qualitatively distinct responses, no study has compared them directly.

The majority of unanticipated question studies comprise post-hoc analyses of interviewees' verbal behaviour, looking at differences in the level of detail, consistency and statement length (Mac Giolla & Granhag, 2015; Sooniste et al., 2015; Vrij et al., 2009; Warmelink et al., 2012). Vrij, Leal and colleagues (2011, 2012) conducted follow-up studies where observers made veracity judgements from interview transcripts, finding that accuracy was greater than chance only with transcripts containing unanticipated questions. However, no studies have required interviewers to make real-time veracity judgements. The goal of many investigative interviews (e.g., interrogations in the US justice system, security screening, and vetting interviews) is to allow the interviewer to establish the veracity of the interviewee's account. In a study by Sooniste, Granhag, and Strömwall (2017), experienced police officers were trained

to detect deception using, among other methods, unanticipated questions. Subsequently, they interviewed truth-tellers and liars and were required to make real-time veracity judgements. The officers who were trained performed better than untrained officers, though this difference in accuracy was not significant. However, they were given the freedom to conduct the interview as they chose and so it was only possible to measure the presence of unanticipated questions in a post-hoc fashion.

Unanticipated questions may elicit verbal cues to deceit, but their effects on judgements of the interviewer are unknown. Vrij and colleagues (2017) meta-analysis into the cognitive approach to lie detection, which uses unanticipated questions, found across studies that veracity was correctly classified 71% of the time when using this technique, compared with only 56% using standard interview approaches. However, Levine, Blair, and Carpenter (2018) recently challenged these findings, arguing that the meta-analysis confounded dependent variables by combining human veracity detection rates and statistical classifications based on coded differences in interview transcripts. By re-examining the data, they showed a difference in accuracy rates obtained by the two outcome measures, with higher rates observed for statistical classifications (78%) than human judgements (62%). Therefore, it remains unclear whether statistical differences in verbal behaviour translate to an improvement in human veracity judgement accuracy.

The studies presented below examined the effects of unanticipated questions using three different empirical approaches. Experiment 1 provided a within-experiment comparison of the effectiveness of unanticipated planning and unanticipated spatial/temporal questions, to determine if the use of unanticipated questions leads to improved accuracy in the real-time veracity judgements made by interviewers. The resulting interviews were analysed using the Reality Monitoring (RM) framework

(Johnson & Raye, 1981) to examine whether anticipated and unanticipated questions generate differences in verbal content of truth-tellers' and liars' responses. In Experiment 2, transcripts of the interviews conducted in Experiment 1 were shown to a separate group of observers, who were required to make a veracity judgement.

Experiment 1

In this experiment, truth-tellers completed a task which involved navigating around a university campus, while liars had to pretend to have conducted the same task. All interviewees were subsequently told to convince an interviewer that they had carried out the task. The interview questions were either questions that might be anticipated by interviewees (e.g., "What task did you carry out around the campus today?"), unanticipated questions about the planning of the task (e.g., "Please describe any changes you made to your plan during the planning stage"), or unanticipated questions regarding spatial and temporal details (e.g., "In building B, where were the boxes in relation to the door you entered through?"). The planning questions were based on questions asked in the experiments conducted by Sooniste and colleagues (2013) and Granhag and colleagues (2016), while the spatial and temporal questions were based on those used by Vrij and colleagues (2009) and Lancaster and colleagues (2013). Immediately following the interviews, interviewers made a veracity judgement concerning the interviewee's account and were asked what information they based their decision on.

Based on previous work by Vrij and colleagues showing unanticipated questions in interviews results in differences in the verbal behaviour of truth-tellers and liars (e.g., Lancaster et al., 2013; Warmelink et al., 2012), interviewers should make more accurate veracity judgements when asking questions regarding planning or spatial and temporal

details that are unlikely to be anticipated by interviewees than when asking the general questions about the event that are likely to be anticipated (Hypothesis 1). The unanticipated questions approach is grounded in the idea that liars will experience an increase in cognitive load when answering unanticipated questions compared to ones they have anticipated, while truth-tellers should experience similar levels across question type (Vrij, 2014). As such, liars should give higher ratings of cognitive complexity to the interviews involving unanticipated questions than the anticipated questions, with no such differences observed between the ratings given by truth-tellers (Hypothesis 2). Finally, given that the unanticipated questions approach is said to elicit differences in the verbal content of truth-tellers' and liars' accounts (Vrij, 2014), interviewers who reported verbal content as the basis for their decisions should show greater judgement accuracy (Hypothesis 3). A failure to find support for each of these hypotheses would cast doubt upon the unanticipated questions framework.

The experiment also investigated differences in the verbal responses provided by truth tellers and liars, and whether they are amplified by asking unanticipated questions. The RM framework (Johnson & Raye, 1981) asserts that an individual's memory of a genuine experience is intrinsically associated with perceptual processes, meaning they will be richer in details related to sensory information (e.g., visual and auditory), contextual information (e.g., spatial and temporal) and affective information (e.g., references to emotional state; Vrij, Mann, Kristen, & Fisher, 2007). Accounts of imagined experiences are conceived endogenously, without any genuine perceptual information, meaning they are likely to be richer than accounts of genuine experiences in cognitive operations (e.g., references to thought processes; Oberlader et al., 2016). RM has been utilised in deception research, with several studies reporting it can

distinguish between true and false accounts (Logue, Book, Frosina, Huizinga, & Amos, 2015; Memon, Fraser, Colwell, Odinot, & Mastroberardino, 2010; Vrij et al., 2007).

Unanticipated questions are designed to force the interviewee into providing a spontaneous, unprepared answer and as such a dishonest interviewee should have less opportunity to access related experience from memory (Vrij, 2014). Research has shown that unanticipated questions emphasise differences in truth-tellers' and liars' verbal behaviour in terms of statement length and level of detail (Sooniste et al., 2013). These amplified differences should be detected by RM. Although there has been variation among studies that have utilised RM in terms of the linguistic categories used, the four most commonly associated with deception are words relating to sensory information (e.g. "saw", "heard"), contextual information (e.g. "up", "after"), affective information (e.g. "upset", "pleased"), and cognitive mechanisms (e.g. "thought", "considered"). Previous research has shown that truth tellers tend to use more sensory and contextual information words than liars (Memon et al., 2010) given that they have a true episodic memory of the event in question, which is likely to be rich in perceptual information (Vrij et al., 2007). Liars, on the other hand, have been shown to use more words related to cognitive mechanisms than truth tellers (Logue et al., 2015) because they must rely on imagined experience of the event, without genuine perceptual information (Oberlader et al., 2016). Research on the affective information category is less clear. The original theory on which RM is based states that truth-tellers should use more affective information words than liars (Johnson & Raye, 1981), and this pattern has previously been reported (Sporer, 1997). However, some findings show no difference between truth tellers and liars (Logue et al., 2015; Memon et al., 2010).

The number of words falling into the four RM categories was measured for each interview transcript using the linguistic analysis software LIWC (Pennebaker, Booth,

Boyd, & Francis, 2015). Based on RM theory (Johnson & Raye, 1981) and previous findings specific to deception (Logue et al., 2015; Memon et al., 2010; Sporer, 1997), truth tellers should use more words associated with sensory, contextual and affective information and liars should use more words associated with cognitive mechanisms than truth tellers (Hypothesis 4). Additionally, based on the findings of Vrij and colleagues regarding the unanticipated questions approach (Lancaster et al., 2013; Vrij et al., 2009), differences in the verbal content of truth tellers' and liars' responses should be amplified by the use of unanticipated questions (Hypothesis 5).

Method

Participants

Interviewees. Sixty interviewees were assigned to the truth-teller condition. Of these, 42 were female (M_{age} = 21.52, SD = 4.32) and 18 were male (M_{age} = 23.00, SD = 6.38). A further 60 interviewees were assigned to the liar condition. Of these, 47 were female (M_{age} = 20.38, SD = 2.65) and 13 were male (M_{age} = 22.69, SD = 4.23). Interviewees were UG and PG students recruited from a range of science and arts disciplines at the University of Sussex. Interviewees received either course credits or £5 for taking part. As an additional incentive, they were told that they would receive a further £5 if they were successful in convincing the interviewer that they were telling the truth. In reality, all interviewees received this extra money regardless of performance. This study was approved by the Sciences & Technology Cross-Schools Research Ethics Committee at the University of Sussex. All participants provided written consent.

Interviewers. Six female (M_{age} = 29.67, SD = 5.09) and four male (M_{age} = 30.75, SD = 10.91) Psychology doctoral students at the University of Sussex were

selected to carry out the interviews. All attended training which comprised classroom-based instruction and practical exercises on using the interview protocol devised for this research, which consisted of a fixed set of questions varying by condition (Appendix 1). During the training exercise, they were informed about the importance of sticking to the protocol, asking all questions on the question list they had been provided with, and to avoid asking additional follow-up questions. They were also given practical advice regarding methods for detecting deceit (e.g., the importance of carefully monitoring the interviewee's verbal behaviour). Finally, they each were required to conduct a practice interview with the experimenter, in order to ensure that they understood the procedure. Interviewers were given basic information about the task that the interviewees were going to be carrying out, but all were blind to the veracity of the interviewees and hypotheses of the study. Each conducted twelve interviews and was paid £65 for taking part.

Design

A between-groups design was employed, with interviewees randomly assigned to either truth-teller (n = 60) or liar (n = 60) conditions. Interviewees were further randomly assigned to one of three interview conditions: anticipated (n = 40), planning (n = 40), or spatial/temporal (n = 40). Assignment was balanced across condition so that for each of the interview conditions, half were truth-tellers and half were liars.

Procedure

Truth-tellers. Those assigned to the truth-teller condition arrived at the interview room and, after reading an information sheet and signing a consent form, were escorted to a room in another building on campus, where they received written instructions. The instructions informed the participant that they were currently in Room

A and that in front of them they would see a stack of paper box files, each a different colour. In each trial, the number of boxes left in Room A was varied between two and four in order to prevent the interviewers being able to determine veracity based on the number of boxes left at Room A. The goal was to ensure that there were five boxes stacked in Room A by the end of the task, so interviewees should collect further boxes from Room B, located within another building on a remote side of the campus that is not frequented by anyone other than maintenance staff. They were also informed that the entrance to Room B had an access code and that, although one of the experimenters should be there to let them in, they should consider alternative routes in case the experimenter was unable to be there. In reality, the experimenter was always there to let them in. This instruction was included in order to create a scenario which would require a degree of forward planning by interviewees, and to introduce a degree of ambiguity to prevent interviewers from learning task-induced differences between truth-teller and liar accounts. They were instructed to take five minutes to plan how they would complete the task and then no more than 30 minutes to complete it and then return to the interview room. In order to encourage interviewees to spend time planning the task, the instructions again stated that they should consider both the time limit and the possibility that they would be unable to enter Room B via the main entrance. Prior to the interview following the task, they were instructed to answer all questions as accurately and honestly as possible. Interviewees were given a campus map that highlighted Room A and Room B. Interviewees kept track of time using their watch or phone.

Liars. Liars were informed that they would not be carrying out the navigation task but that their goal was to convince the interviewer that they had, and that they would have to answer interview questions dishonestly. They were given instructions for

the navigation task, which were the same as the instructions given to truth-tellers, including the information regarding the potential complications accessing Room B, and the map of the campus. They were given five minutes planning time to develop a convincing story that would help them answer the interviewer's questions.

The interview. Prior to each interview, the interviewer was handed one of three question lists and then was introduced to the interviewee. The experimenter turned on the two cameras (one directed at the interviewee and one at the interviewer) and then left the room, leaving the interviewer to ask the set of questions. Each of the question lists consisted of ten questions. The first five questions were the same in each list and consisted of general questions about the task that interviewees might have anticipated, such as "What task did you carry out around the campus today?" and "Describe the route you took from Room A to Room B." The remaining questions differed according to condition: In the general condition, they were further general questions similar to the first five, such as "How many boxes were there in Room B?" In the planning condition, questions (adapted from those asked in both Sooniste et al., 2013 and Granhag et al., 2016) focussed on the planning of the task, such as "Explain what steps you would have taken had you not been able to access Room B via the main door" and "Please describe any changes you made to your plan during the planning stage." In the spatial/temporal condition, the questions (adapted from both Vrij, et al., 2009 and Lancaster, et al., 2013) focussed on spatial and temporal details, such as "Try to imagine the layout and features of the Room B. Please describe this room and be as detailed as you can" and "Please describe the task in full, but now in reverse order."

In order to prevent the interviewers from gaining advantageous information as the experiment progressed, (e.g., that an experimenter was always in place at Room B),

they were never given feedback on their performance until all twelve interviews had been completed.

Post-interview questionnaires. When the interview was complete, the interviewee completed two questionnaires (Appendix 3 & 4). The first listed the ten questions that they had been asked and required them to state, using a 7-point Likert scale, how much they had anticipated each question prior to interview. The second gathered information, again using 7-point Likert scales, regarding how deceptive/truthful they had been, how cognitively demanding they found the interview, and how motivated they were to comply with the instructions.

The interviewers also completed a questionnaire (Appendix 2) after each interview in which they indicated whether they felt the interviewee had been lying or telling the truth, firstly on a 7-point Likert scale and secondly using a dichotomous forced choice decision. The questionnaire also required them to explain any verbal or non-verbal information they had based their decision on. Responses were subsequently coded as one of four categories: Verbal Content, such as "specific details in responses" or "consistency in responses"; Verbal Delivery, such as "tone of voice" or "responses seeming rehearsed"; Non-verbal Passive, such as "nervous demeanour" or "level of confidence"; and Non-verbal Active, such as "hand movements", "body language" or "eye contact".

Results

Manipulation Checks

Interviewee compliance. Interviewees were asked to rate the extent to which they had been deceptive in the interview on a seven-point Likert scale (1 = totally truthful; 7 = totally deceptive). Interviewees assigned to liar conditions reported being

more deceptive (M = 6.27, SD = 0.86) than those in the truth-teller condition (M = 1.15, SD = 0.71), t (118) = -35.54, $p < .001$, d = 6.49, 95% CI [5.56, 7.34]. Motivation to comply was high in both groups, with no difference in ratings between truth tellers (M = 6.08, SD = 1.05) and liars (M = 6.10, SD = 0.86), t (118) = -0.10, p = .93.

Interviewer compliance. Transcripts of the interviews were assessed to establish whether the interviewers had adhered to the interview protocol. The total number of deviations from the 10-question script was calculated for each interview. Deviations included missing out a question, changing the order of the questions, altering the wording of a question, asking an incomplete question, or asking an additional question. Overall, the number of deviations in each interview was low (M = 0.80, SD = 1.12) and the majority were due to interviewers slightly rephrasing questions to help the interviewee understand. A 2 (veracity: truth-teller or liar) × 3 (question type: anticipated, unanticipated planning, or unanticipated spatial/temporal) between-groups ANOVA showed no main effect of veracity, F (1, 114) = 2.13, p = .15, nor a main effect of question type, F (1, 114) = 0.41, p = .66. There was also not a significant veracity × question type interaction, F (1, 114) = 0.38, p = .69.

Anticipation. Interviewees rated the extent to which they had anticipated each question on a seven-point scale (1 = completely expected; 7 = completely unexpected). Mean anticipation was calculated for the final five questions of each interview. A 2 (veracity: truth-teller or liar) × 3 (question type: anticipated, unanticipated planning, unanticipated spatial/temporal) between-groups ANOVA showed no main effect of veracity, F (1, 114) = 0.95, p = .33, nor a significant veracity × question type interaction, F (2, 114) = 1.45, p = .24. There was a significant main effect of question type, F (2, 114) = 20.83, $p < .001$, $\eta_p^2 = .27$, 95% CI [.13, .38]. Planned contrasts revealed that questions assigned to the anticipated conditions (M = 4.11, SD = 1.24)

were significantly more anticipated than questions assigned to the unanticipated conditions (i.e., the average of planning and spatial/temporal questions combined; M = 5.36, SD = 0.87), $F = 41.30$, $p < .001$, $d = 1.24$, 95% CI [0.82, 1.64] . However, there was no significant difference in anticipation of planning questions (M = 5.43, SD = 0.85) and spatial/temporal questions (M = 5.30, SD = 0.89), $F = 0.36$, p = .55.

Accuracy

Forced choice. The interviewer made a dichotomous decision post-interview regarding the veracity of each interviewee and did so for two interviewees in each of the six conditions. The overall mean accuracy was 54%. A one-sample *t*-test showed that this was not significantly different from chance (50% correct), t (119) = 0.91, $p = .36$. In a series of one-sample t-tests (see Table 1) accuracy when asking anticipated questions was significantly better than chance at identifying truth-tellers, t (19) = 2.52, $p = .021$, $d = 0.56$, 95% CI [0.08, 1.03]. However, performance was significantly worse than chance at identifying liars, t (19) = -3.27, $p = .004$, $d = 0.73$, 95% CI [0.23, 1.22]. For truth-tellers and liars combined, performance was not significantly different from chance, t (39) = -0.31, $p = .76$.With unanticipated planning questions, performance did not differ from chance at interviewing truth-tellers, t (19) = 0.89, $p = .39$, liars, t (19) = -0.89, $p = .39$, or for truth-tellers and liars combined, t (39) = 0, $p = 1$. With unanticipated spatial/temporal questions, interviewers were significantly better than chance at identifying truth-tellers, t (19) = 2.52, $p = .021$, $d = 0.56$, 95% CI [0.08, 1.03], but not at identifying liars, t (19) = 0.44, $p = .67$. For truth-tellers and liars combined, interviewer accuracy with the unanticipated spatial/temporal questions was also not significantly greater than chance, t (39) = 1.96, $p = .06$.

Table 1.

Mean (SD) accuracy rates across each question type for both truth-tellers, liars, and overall.

Question Type	Truth-teller	Liar	Overall
Anticipated	**75%** (44%)	**20%** (41%)	48% (51%)
Planning	60% (50%)	40% (50%)	50% (51%)
Spatial/Temporal	**75%** (44%)	55% (51%)	65% (48%)

Note: Bold figures indicate that the accuracy significantly differed from chance (50%)

To investigate the relative effects of veracity and question type on the interviewers' dichotomous judgement accuracy (where scores varied between 0 and 2, interviewers contributing two judgements in each condition), a 2 (veracity: truth-teller or liar) × 3 (question type: anticipated, unanticipated planning, or unanticipated spatial/temporal) repeated measures ANOVA was conducted. There was a significant effect of Veracity, F (1, 9) = 13.05, p = .006, η_p^2 = .59, 95% CI [.08, .77], with overall accuracy greater for truth-tellers (70%) than for liars (38%). Neither Question Type, F (2, 8) = 1.56, p = .27, nor the interaction between Veracity and Question Type, F (2, 8) = 2.45, p = .15, was significant.

Veracity scale. Interviewers were also asked to rate the extent to which they thought the interviewee was telling the truth or lying on a seven-point scale (1 = definitely lying; 7 = definitely telling the truth). Scores in the liar conditions were reversed so that higher scores indicate greater accuracy. Figure 1 shows the mean scores given across the three interview types for truth-tellers and liars.

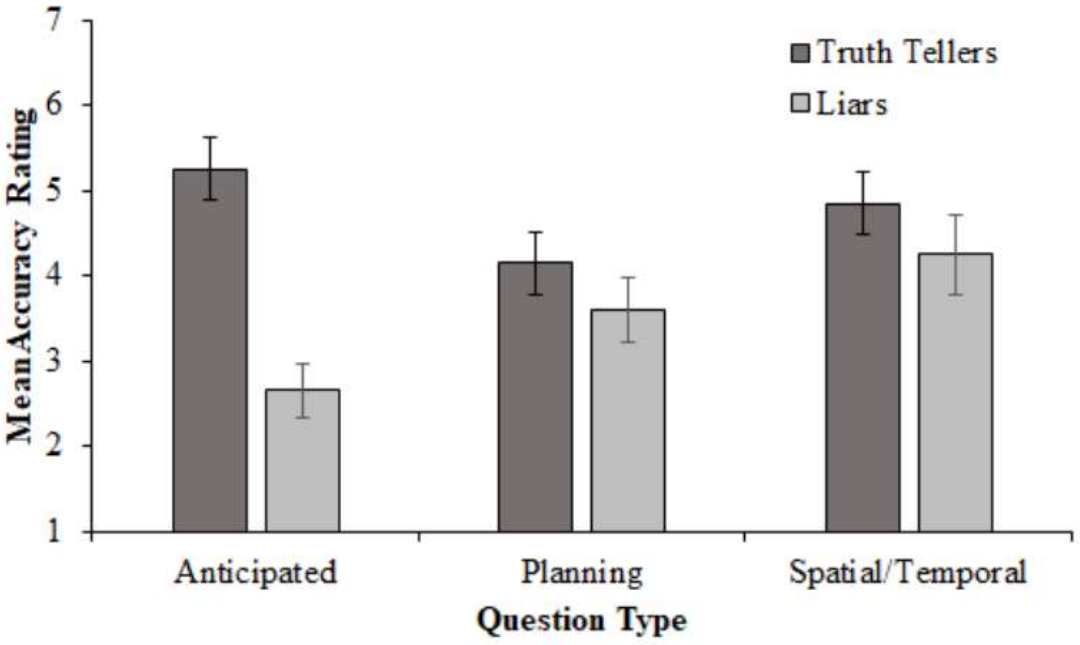

Figure 1. Mean interviewer accuracy (measured via 7-point Likert scale) across question type for truth-tellers and liars separately. Error bars represent +/- 1 SEM.

To investigate the effects of veracity and question type, a 2 (veracity: truth-teller or liar) × 3 (question type: anticipated, unanticipated planning, or unanticipated spatial/temporal) between-groups ANOVA was performed on level of accuracy. There was a significant main effect of veracity, F (1, 114) = 16.10, p = < .001, η_p^2 = .12, 95% CI [.03, .24], as well as a significant interaction, F (2, 114) = 4.70, p =.011, η_p^2 = .08, 95% CI [.004, .17]. There was no main effect of question type, F (2, 114) = 1.88, p = .157. Planned contrasts reveal that accuracy was greater for truth-tellers (M = 4.75, SD = 1.67) than for liars (M = 3.50, SD = 1.86), F = 16.10, p < .001, d = 0.71, 95% CI [0.32, 0.93]. The difference in accuracy between truth-tellers and liars was significantly greater for the anticipated questions (M_{diff} = 2.60, SD = 3.08) than the two unanticipated question types combined (M_{diff} = 0.56, SD = 3.59), F = 9.36, p = .003, d = 0.61. However, there was no difference between the unanticipated spatial/temporal and unanticipated planning questions, F = 0.01, p = .95.

Cognitive Demand

Interviewees were asked to rate how cognitively demanding they found the interview on a seven-point scale (1 = very easy; 7 = very difficult). Figure 2 shows the mean ratings given to each question type for truth-tellers and liars. To investigate the effects of veracity and question type, a 2 (veracity: truth-teller or liar) × 3 (question type: anticipated, unanticipated planning, unanticipated spatial/temporal) between-groups ANOVA was conducted on cognitive demand ratings. There were main effects of both veracity, F (1, 114) = 95.32, $p < .001$, $\eta_p^2 = .46$, 95% CI [.32, .56] and question type, F (2, 114) = 13.75, $p < .001$, $\eta_p^2 = .19$, 95% CI [.07, .31]. However, there was no significant interaction, F (2, 114) = 0.02, $p = .98$. Planned comparisons revealed that, overall, liars found the interviews more difficult (M = 4.47, SD = 1.49) than the truth-tellers (M = 2.32, SD = 1.13), $F = 95.32$, $p < .001$, $d = 1.63$, 95% CI [1.20, 2.03]. Interviewees found the unanticipated questions combined (M = 3.71, SD = 1.66) more cognitively demanding than the anticipated questions (M = 2.75, SD = 1.61), $F = 16.98$, $p < .001$, $d = 0.58$, 95% CI [0.19, 0.97]. Additionally, spatial/temporal questions (M = 4.15, SD = 1.70) were rated as significantly more cognitively demanding than planning questions (M = 3.28, SD = 1.52), $F = 10.53$, $p = .002$, $d = 0.53$, 95% CI [0.09, 0.98].

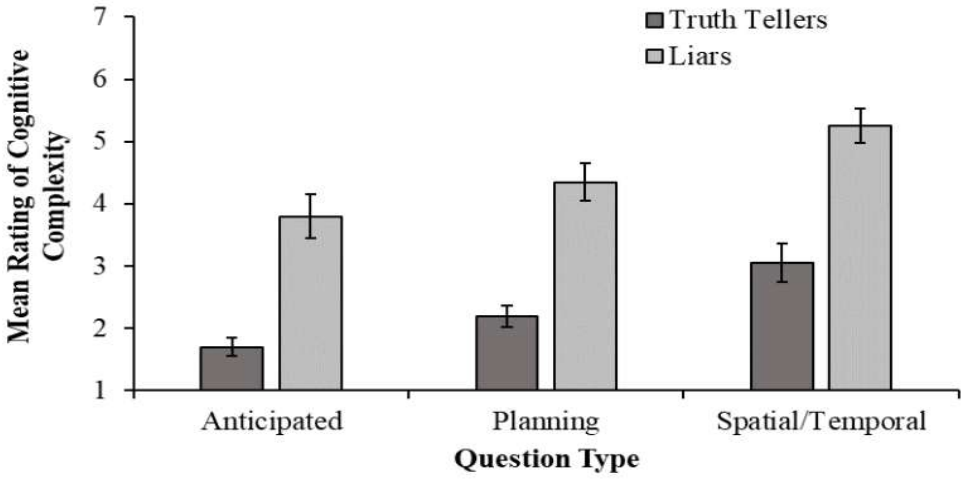

Figure 2. Mean ratings of cognitive complexity (1 = very easy, 7 = very difficult) given to each interview type for truth-tellers and liars separately. Error bars represent +/-1 SEM.

Perceived Cues

The reasons that the interviewers reported for their veracity decisions were grouped into four categories: verbal content, verbal delivery, non-verbal passive, and non-verbal active. The total number within each category was calculated for each interview. A multiple regression was performed using these frequencies as predictors with accuracy (judged via the 7-point veracity scale) as the dependent variable. Verbal content was entered alone in the first step of the model, given that the unanticipated questions approach relies on detecting differences in the verbal content of interviewees' responses (Vrij, 2014), and verbal delivery, non-verbal passive and non-verbal active were entered together at the second step (see Table 2). The model at Step 1 was significantly able to predict interviewer accuracy, F (1, 118) = 6.22, p = .01, η_p^2= .05, 95% CI [.002, .14]. The greater the number of verbal content reasons that interviewers claimed to base their decisions on, the greater their accuracy level was. This provides support for Hypothesis 3. Despite this, the model was only able to explain 5% of the variance in accuracy. The model remained significant at Step 2, F (4, 115) = 2.89, p =

.03, η_p^2 = .09, 95% CI [.001, .17], however, the addition of the three remaining predictors did not significantly improve the model, $\Delta R^2 = .04$, F (3, 115) = 1.74, $p = .16$. Inspection of the data for Step 2 reveals that verbal delivery, and both non-verbal categories were negatively related to interviewer accuracy, indicating that the more of these types of reasons that interviewers based their decisions on, the worse their accuracy became. However, none of these were significant predictors.

The analysis was repeated, with a binary logistic regression, using forced choice accuracy as the dependent variable. The findings were essentially the same as those of the Likert scale data. The model at Step 1 was significantly able to predict interviewer accuracy, $\chi^2(1) = 4.07$, $p = .05$. The greater the number of verbal content reasons that interviewers claimed to base their decisions on, the greater their accuracy levels. Despite this, the model was only able to explain 5% of the variance in accuracy (Nagelkerke R^2). The model was no longer significant at step two. The addition of the three remaining predictors did not significantly improve the model, $\chi^2(3) = 3.92$, p = .27.

Table 2

Regression outcome for post-interview veracity decision (truth-teller versus liar) made by interviewers

				95% CI for *b*	
	b	SE *b*	*β*	Lower	Upper
Step 1					
Constant	3.49	0.31		2.89	4.09
Verbal Content	0.50	0.20	.22*	0.10	0.90
Step 2					
Constant	4.25	0.45		3.36	5.14
Verbal Content	0.38	0.21	.18	-0.02	0.79
Verbal Delivery	-0.38	0.28	-.13	-0.92	0.17
Non-Verbal Passive	-0.24	0.27	-.08	-0.77	0.30
Non-Verbal Active	-0.23	0.14	-.15	-0.51	0.05

Note: R^2 = .05 for Step 1 (p = .01), ΔR^2 = .04 for Step 2 (p = .16). * p < .05.

Reality Monitoring Analysis

Analysis approach. The text analysis software programme Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015) was used to carry out word counts in this study. In order to investigate the effects of veracity and question type, only transcripts of the final five questions in each interview were included in analysis (the first five being common to all conditions). To prepare the transcripts for analysis, all utterances from the interviewer were removed, leaving only responses made by interviewees. Responses from each interview (including utterances, such as ‘er’ or ‘hmm’) were entered together as one paragraph. Filler words, such as ‘you know’, were transcribed as one word (e.g., ‘youknow’). Finally, the word ‘like’, when used as a filler word, was transcribed as ‘rrlike’ in order to be classified as such by LIWC.

For each transcript, LIWC determines the amount of words falling into 73 linguistic categories, each presented as percentages of total word count. Four of

relevance to RM were analysed: ‘perceptual processes’, ‘relativity’, ‘affective processes’, and ‘cognitive mechanisms’. The ‘perceptual processes’ (or ‘sensory’) category includes words relating to sensory information, such as ‘saw’, ‘heard’, and ‘felt’. The ‘relativity’ (or ‘contextual’) category includes spatial and temporal related words, such as ‘down’, ‘arrive’, and ‘during’. The ‘affective processes’ category includes emotion-based words, both positive and negative, such as ‘happy’, ‘hurt’, and ‘worried’. Finally, the ‘cognitive mechanisms’ category includes words associated with thought processes, such as ‘know’, ‘think’, ‘maybe’ and ‘because’. These categories are similar to those used by Bond and Lee (2005).

Word count. In order to explore the effects of veracity and question type on the total number of words spoken by interviewees, a 2 (Veracity: truth-teller or liar) × 3 (Question Type: anticipated, unanticipated planning, or unanticipated spatial/temporal) ANOVA was conducted with word count as the dependent variable (Figure 3). There was no effect of veracity, F (1, 114) = 0.45, p = .50, nor was there a significant interaction, F (2, 114) = 0.81, p = .45. However, there was a significant main effect of question type, F (2, 114) = 7.52, p = .001, η_p^2 = .12, 95% CI [.02, .22]. Post-hoc tests revealed a significantly lower word count in response to unanticipated planning questions (M = 190.88, SD = 115.54) than to both anticipated questions (M = 285.98, SD = 130.62), t (78) = -3.45, p = .001, d = 0.77, 95% CI [0.31, 1.22] and unanticipated spatial/temporal questions (M = 293.23, SD = 145.81), t (78) = -3.48, p = .001, d = 0.78, 95% CI [0.32, 1.23]. There was no significant difference in word count between responses to anticipated questions and unanticipated spatial/temporal questions, t (78) = -0.23, p = .82.

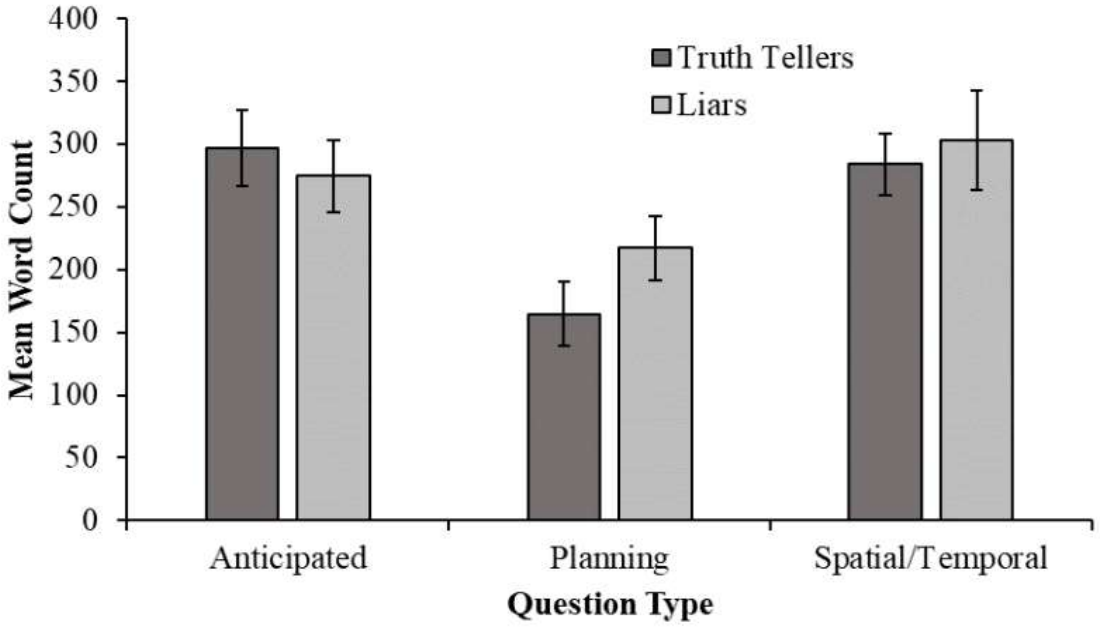

Figure 3. Mean word count of the responses provided by interviewees. Error bars represent +/-1 SEM

Category counts. Table 3 shows the mean percentage of truth-tellers' and liars' statements in each RM category for each of the three question types. To examine the effects of veracity and question type, a 2 (Veracity: truth-teller or liar) × 3 (Question Type: anticipated, unanticipated planning, or unanticipated spatial/temporal) MANOVA was conducted with the four RM criteria entered as dependent variables. There were significant overall main effects of veracity, F (4, 111) = 2.59, p = .04, η_p^2 = .09, 95% CI [.001, .17], and question type, F (8, 224) = 16.10, p < .001, η_p^2 = .37, 95% CI [.25, .43]. Additionally, there was a significant veracity × question type interaction, F (8, 224) = 2.01, p = .046, η_p^2 = .07, 95% CI [.001, .11].

Table 3.

RM category mean (SD) counts for each question type.

	Truth tellers	Liars
Anticipated Questions		
Sensory	2.30 (1.38)	1.51 (0.96)
Contextual	20.77 (3.60)	19.68 (4.03)
Affective	2.09 (0.86)	2.25 (1.25)
Cognitive Mechanisms	8.82 (2.22)	8.94 (2.01)
Planning Questions		
Sensory	1.55 (1.83)	1.93 (1.15)
Contextual	17.20 (5.41)	18.01 (3.64)
Affective	1.89 (1.27)	3.07 (1.27)
Cognitive Mechanisms	16.28 (4.38)	17.80 (3.30)
Spatial and Temporal Questions		
Sensory	1.90 (1.00)	1.32 (0.57)
Contextual	22.64 (5.26)	23.99 (5.60)
Affective	1.59 (0.87)	1.92 (1.05)
Cognitive Mechanisms	8.53 (4.02)	7.96 (2.90)

Univariate tests of the four RM criteria revealed a significant effect of veracity with affective words, F (1, 114) = 7.59, p = 0.01, η_p^2 = .06, 95% CI [.01, .16], showing that liars (M = 2.41, SD = 1.27) used significantly more affective words than truth tellers (M = 1.86, SD = 1.01), t (118) = -2.65, p = .01, d = 0.48, 95% CI [0.12, 0.85]. The effects of veracity with the remaining RM criteria were not significant (all ps > .05).

There was a significant univariate effect of question type on contextual words, F (2, 114) = 15.00, p < .001, η_p^2 = .21, 95% CI [.08, .32], with significantly more in response to unanticipated spatial/temporal questions (M = 23.32, SD = 5.40) than to both anticipated questions (M = 20.22, SD = 3.81), t (78) = 2.96, p = .004, d = 0.66, 95% CI [0.21, 1.11], and unanticipated planning questions (M = 17.61, SD = 4.57), t

(78) = 5.10, $p < .001$, $d = 1.14$, 95% CI [0.66, 1.61]. There was also an effect of question type on affective words, F (2, 114) = 4.33, $p = .02$, $\eta_p^2 = .07$, 95% CI [.003, .16], with significantly more in response to unanticipated planning questions (M = 2.48, SD = 1.39) than to unanticipated spatial/temporal questions (M = 1.75, SD = 0.97), t (78) = 2.72, $p = .01$, $d = 0.61$, 95% CI [0.16, 1.05]. Finally, there was a significant effect of question type on cognitive mechanism words, F (2, 114) = 90.67, $p < .001$, $\eta_p^2 = .61$, 95% CI [.50, .69], with significantly more in response to unanticipated planning questions (M = 17.04, SD = 3.90) than to both anticipated questions (M = 8.88, SD = 2.10), t (78) = 11.65, $p < .001$, $d = 2.61$, 95% CI [2.00, 3.20], and to unanticipated spatial/temporal questions (M = 8.24, SD = 3.47), t (78) = 10.64, $p < .001$, $d = 2.38$, 95% CI [1.80, 2.95]. There was no significant effect of question type on perceptual details, F (2, 114) = 0.60, $p = .55$.

Discussion

The results of Experiment 1 indicate that the manipulations were successful. The planning and spatial/temporal questions were rated as significantly less anticipated than the anticipated questions. Additionally, participants appeared to comply with the instructions and were motivated to do so. As with all subjective response measures, responses to the post-interview questionnaire may have been influenced by study demand characteristics. Nonetheless, the absence of differences between conditions gives us some degree of confidence that the motivation to conform was high and equivalent across conditions. Overall, the findings of Experiment 1 indicate that unanticipated questions did not increase interviewers' ability to detect interviewee veracity. The veracity scale judgements and forced choice results show the same pattern: while accuracy for detecting liars increased slightly with unanticipated

questions, accuracy at detecting truth-tellers was reduced, particularly with planning questions. As such, the findings fail to support Hypothesis 1. Furthermore, the results do not support the idea that unanticipated questions raise cognitive load for liars but not for truth tellers, failing to support Hypothesis 2. The unanticipated questions approach is grounded in the idea that being asked unanticipated questions in an interview will raise the cognitive load for liars but not truth tellers (Vrij, 2014). However, in the present study, liars found the interviews more difficult than truth tellers regardless of question type, and all interviewees found the unanticipated spatial/temporal interviews more cognitively demanding than the anticipated or unanticipated planning interviews, regardless of veracity condition. There was, however, a small positive correlation between accuracy and the number of verbal content reasons interviewers claimed to base their veracity judgements on, supporting Hypothesis 3.

Previous research has shown that truth tellers use more words associated with sensory, contextual and affective processes than liars, while liars tend to use more cognitive mechanism words than truth tellers (Logue et al., 2015; Memon et al., 2010; Sporer, 1997). The present study found a difference in the number of affective words given by liars and truth tellers, providing modest support for Hypothesis 4. Truth-tellers and liars used qualitatively different language in response to the three question types, with more contextual detail words when answering the spatial/temporal questions and more cognitive mechanism words with planning questions. However, although a significant interaction was found between veracity and question type, at a univariate level there was no significant effect for any of the four RM categories, thus Hypothesis 5 was rejected. It seems that the content of unanticipated questions alters the content of answers, but not in a way that discriminates truth-tellers from liars.

Experiment 2

For tasks such as security screening and police stop-and-search interviews, methods are needed that can be used to determine interviewee veracity in real time. However, in other contexts, the ability to detect deception in a post-hoc fashion is also important. For example, UK police officers are trained according to the PEACE model of investigative interviewing, which states that the purpose of such interviews is to gather information for use by others rather than to determine guilt or innocence directly (Soukara, Bull, Vrij, Turner, & Cherryman, 2009). The information gathered by interviewers, including interview transcripts, may then be used by independent observers, such as judges and juries, in subsequent legal proceedings. Therefore, in Experiment 2, transcripts of the interviews gathered in Experiment 1 were shown to a group of observers who were required to make veracity judgements.

Experiment 1 found that interviewees used qualitatively different language in response to the three question types, with planning questions yielding more references to cognitive operations and spatial/temporal questions yielding more contextual words. Experiment 1 failed to support the UQ approach in terms of its ability to allow interviewers to accurately determine the interviewees' veracity. However, there was a positive relationship between interviewers' reported reliance on verbal content when making veracity judgements and their accuracy. Despite this, the literature on detecting deception suggests that individuals rarely base decisions purely on verbal cues, and instead tend to focus on non-verbal behaviour such as eye contact, body movements, and general demeanour (Colwell et al., 2006; Global Deception Research Team, 2006). The interviewers in Experiment 1 often reported using such non-verbal indicators when making veracity judgements. As such, it is possible that poor accuracy rates could be attributed to interviewers relying on ineffective non-verbal cues (DePaulo et al., 2003),

as opposed to more useful verbal cues elicited by unanticipated questions. Experiment 2 was conducted in order to determine whether unanticipated questions could improve veracity judgement accuracy when non-verbal behaviour is not present to influence decision making. Previous observer studies have reported positive results. For example, Vrij, Leal and colleagues (2011, 2012) found that observers were able to accurately determine the veracity of interviewees when the transcripts contained unanticipated questions, but not from transcripts containing only anticipated questions.

Based on these findings (Vrij, Leal, et al., 2011, 2012), as well as research into the unanticipated questions approach showing differences between truth-tellers' and liars' verbal behaviour (Lancaster et al., 2013; Vrij et al., 2009; Warmelink et al., 2012), we expected to find that observers would show higher levels of accuracy when judging the veracity of transcripts containing unanticipated questions, compared to those containing anticipated questions (Hypothesis 6).

Method

Participants

Ninety females (M_{age} = 30.30, SD = 16.40) and 21 males (M_{age} = 34.62, SD = 17.78) took part in the study. The participants were prospective university students and their parents who voluntarily took part in the experiment as part of an Open Day at the University of Sussex. All gave their informed consent to take part and were free to withdraw at any point. This study was approved by the Sciences & Technology Cross-Schools Research Ethics Committee at the University of Sussex.

Design

A repeated measures design was employed. There were three different interview question types (anticipated, unanticipated planning, and unanticipated spatial/temporal),

each answered by either a truth-teller or a liar, creating a total of six conditions. Each participant was presented with one randomly selected transcript from each of the six conditions.

Procedure

Transcripts were taken from the interviews which took place during Experiment 1. Experiment 2 used transcripts of the final five questions in each interview. In order to moderate effects of variation in interviewee response length, the number of words used by the interviewee in each interview was analysed and the lowest and highest five in each of the six conditions were excluded, leaving ten transcripts per condition (see Table 4 for means).

Table 4.

Mean (SD) word count of transcripts in each condition

Question Type	Truth teller	Liar
Anticipated	268.10 (48.15)	260.80 (46.14)
Planning	154.90 (52.54)	222.40 (72.09)
Spatial and Temporal	263.10 (51.65)	262.70 (96.82)

Participants were informed that they would be reading interview transcripts in which the interviewee may have been telling the truth or lying. They were then told "after reading each transcript, you will be required to state whether you believe the person being interviewed was telling the truth or whether they were lying." Before beginning, the participants were asked to read the instructions for the navigation task that participants received in Experiment 1. Participants were randomly presented on a computer screen with one of ten transcripts from each condition (i.e. they received six transcripts in total) and were given a maximum of three minutes to read each transcript.

The order in which the six conditions appeared on screen was counter-balanced across participants. Following each transcript, they were asked to indicate whether they thought the interviewee was telling the truth or lying via seven-point scale and dichotomous forced choice decision.

Results

Accuracy

Forced choice. Observers made a dichotomous forced choice decision regarding the veracity of the interviewees in each of the transcripts. A series of one-sample t-tests were carried out to investigate effects of veracity and question type on observer accuracy (see Table 5). Looking at detection rates of liars and truth-tellers separately, accuracy at judging anticipated question transcripts was significantly better than chance when identifying truth-tellers, t (110) = 2.22, p = .03, d = 0.21, 95% CI [0.02, 0.40], but not liars, t (110) = -1.43, p = .16. When looking at truth-tellers and liars combined, the observer accuracy rate was not significantly greater than chance, t (110) = 0.55, p = .58.With unanticipated planning transcripts, performance did not significantly differ from chance when identifying truth-tellers, t (110) = 1.63, p = .11, or liars, t (110) = 1.63, p = .11, however, with truth-tellers and liars combined, the accuracy did exceed chance level, t (110) = 2.49, p = .014, d = 0.24, 95% CI [0.05, 0.42]. With unanticipated spatial/temporal transcripts, accuracy levels exceeded chance for both truth-tellers, t (110) = 3.71, p <.001, d = 0.35, 95% CI [0.16, 0.54], and liars, t (110) = 4.65, p < .001, d = 0.44, 95% CI [0.25, 0.64]. When looking at the accuracy rate of truth-tellers and liars combined, observer accuracy was again greater than chance level, t (110) = 5.78, p < .001, d = 0.55, 95% CI [0.35, 0.75].

Table 5.

Mean (SD) observer accuracy rates across each question type for truth-tellers and liars.

Question Type	Truth-teller	Liar	Overall
Anticipated	**60%** (49%)	43% (50%)	52% (34%)
Planning	58% (50%)	58% (50%)	**58%** (32%)
Spatial/Temporal	**67%** (47%)	**70%** (46%)	**68%** (34%)

Note: Bold figures indicate that the accuracy significantly differed from chance (50%)

Veracity scale. As well as making a dichotomous forced choice decision, observers were required to rate whether they thought the interviewee was telling the truth or lying on a seven-point scale (1 = definitely lying; 7 = definitely telling the truth). Scores given to transcripts in the lying condition were reversed meaning that higher scores indicate greater accuracy across all trials. Figure 4 shows the mean scores given across the three question types for truth-tellers and liars.

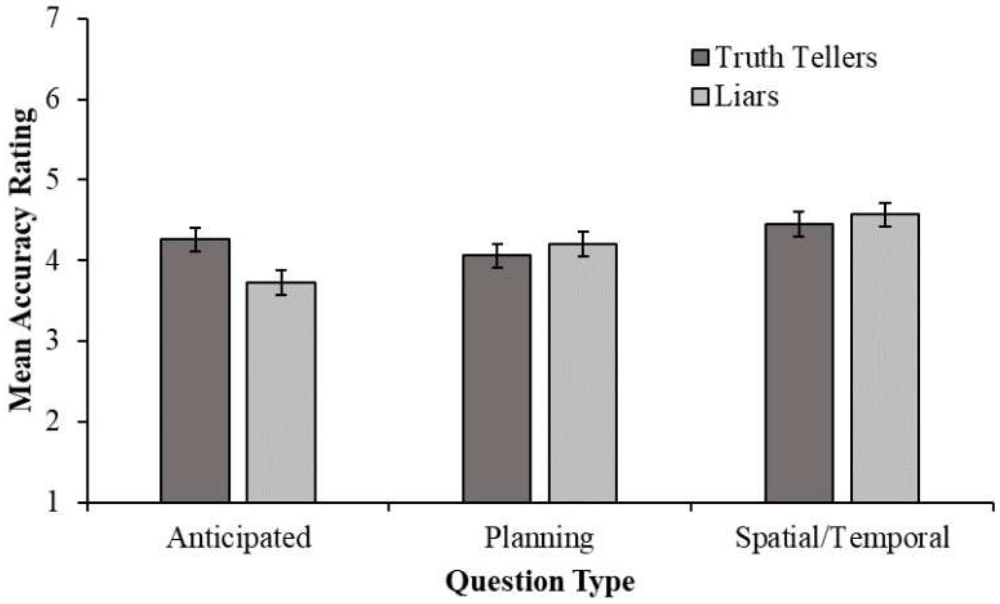

Figure 4. Mean observer accuracy (measured via 7-point Likert scale) across question type for truth-tellers and liars separately. Error bars represent +/- 1 SEM.

A two way 2 (veracity: truth-teller or liar) × 3 (question type: anticipated, unanticipated planning, or unanticipated spatial/temporal) repeated measures ANOVA was performed on rating accuracy. There was no significant main effect of veracity, F (1, 110) = 0.64, p = .43, nor was there a significant interaction, F (2, 220) = 2.88, p =.06. However, there was a significant main effect of question type, F (2, 220) = 6.32, p = .002, η_p^2 = .05, 95% CI [.008, .12]. Planned contrasts revealed that accuracy was significantly greater when observers were judging the transcripts of unanticipated questions (i.e. planning and spatial/temporal questions combined) compared to anticipated questions, F (1, 110) = 6.53, p = .01, η_p^2 = .06, 95% CI [.003, .16]. Furthermore, observer accuracy was significantly higher when judging the spatial/temporal questions than the planning questions, F (1, 110) = 6.13, p = .02, η_p^2 = .05, 95% CI [.002, .15].

Discussion

The findings of Experiment 2 provide only partial support for the unanticipated questions approach (Vrij et al., 2009), and suggest that the type of question asked is crucial. In forced choice judgements, accuracy was greatest when observers were reading transcripts of interviews that included unanticipated spatial/temporal questions. Accuracy when judging transcripts of the anticipated questions was marginally better than chance for truth-tellers, but not liars. When judging the transcripts of planning questions, accuracy was not above chance for truth tellers or liars. When using a scale to make veracity judgements, observer accuracy was greater when judging unanticipated questions than anticipated questions. This is in line with the findings of Vrij, Leal and colleagues (2011, 2012), providing some support for Hypothesis 6, as well as the unanticipated questions approach generally. However, accuracy was also shown to be

higher when observers were judging transcripts of unanticipated spatial/temporal questions compared to transcripts of unanticipated planning questions, which indicates that the type of unanticipated questions asked can have an impact on the ability to determine interviewee veracity.

General discussion

Two experiments explored the effects of different types of unanticipated question on interviewer and observer veracity judgements, and on the content of interviewee accounts. Experiment 1 compared anticipated, unanticipated planning and unanticipated spatial/temporal questions in investigative interviews, with a focus on interviewers' veracity judgement accuracy. The findings fail to provide support for the unanticipated questions approach. With dichotomous forced-choice judgements, accuracy for truth-tellers was no greater when interviewers were asking unanticipated planning or spatial/temporal questions than when asking the anticipated questions. For liars, interviewers were more accurate when asking unanticipated compared to anticipated questions, though neither question type yielded accuracy rates significantly greater than chance. With veracity scale judgements, there was no effect of question type. Accuracy was greater for truth-tellers than liars overall, but this difference was diminished when interviewers asked unanticipated questions compared to the anticipated questions. This suggests that the unanticipated questions approach was marginally useful in improving the detection of liars but impaired the detection of truth-tellers.

According to Vrij (2014), unanticipated questions raise the cognitive load for liars but not for truth-tellers, resulting in observable verbal cues to deceit. In the present study, liars found the interviews more cognitively demanding than truth-tellers. However, all participants found answering unanticipated questions to be more

cognitively demanding than anticipated questions, regardless of veracity condition. This suggests that, while lying is inherently more difficult than telling the truth, the use of unanticipated questions increased the cognitive load faced by liars and truth tellers equally. Previous studies have found similar results, with no interaction between veracity and question type (Sooniste et al., 2013, 2014, 2015). This finding brings into question the proposed underlying mechanism of the UQ approach. Whatever differences there are between truth tellers and liars in their verbal behaviours when answering unanticipated questions, these differences may not be attributable to an increase in cognitive load faced exclusively by liars.

Experiment 1 also revealed that verbal content reasons for veracity decisions were positively associated with judgement accuracy. Verbal content can be a reliable indicator of veracity (DePaulo et al., 2003) and the unanticipated questions approach elicits verbal cues (Vrij et al., 2009). Despite this, the relationship between verbal content and accuracy was small, and the model could only account for 5% of the accuracy variance. Other variables appear to have contributed to accuracy, such as truth bias exhibited in the veracity judgements made by the interviewers. As with all studies of investigative interviews, the extent to which hypothesised base rates of expected truth-tellers and deceivers affected results cannot be assessed. In the present study, interviewers were given no information concerning the base rates for truth-tellers and liars, and this might explain the appearance of a truth bias in interviewer responses. However, the absence of differences between conditions in the presence of truth bias suggests that any impact of underlying base rates was independent of the effects of unanticipated questions. Though, as a result of this bias, accuracy was greater when detecting truth-tellers than liars across all question types, although not at ceiling. The interviewers in Experiment 1 received training. However, none were professionals

within the criminal justice system. Novice veracity judges tend to be biased towards believing an interviewee's account (Granhag & Vrij, 2005; Levine, Park, & McCornack, 1999). It is difficult to control for truth bias. One potential method for future studies would be to inform interviewers in advance that such bias is common. Research into prejudice shows that, by informing an individual of their implicit biases, they are capable of compensating for them (Perry, Murphy, & Dovidio, 2015).

The RM analysis of Experiment 1 found an effect of veracity on affective words, with liars using more than truth-tellers. However, differences in the verbal content of truth tellers' and liars' transcripts were not increased by unanticipated questions. These findings do not support claims that unanticipated questions elicit differences in the verbal behaviour of truth-tellers and liars (Vrij et al., 2009). However, effects of question type were found, with contextual words arising more when answering unanticipated spatial/temporal questions and cognitive mechanism words arising more in responses to unanticipated planning questions. These findings indicate that the type of unanticipated question asked can have a significant effect on the type of information gathered. This may have important implications for determining interviewee veracity. If asking questions about planning taps into an individual's cognitive operations, this may sometimes benefit liars. According to Oberlader and colleagues (2016), liars do not have a genuine perceptual experience of an event to base their responses on and must instead rely on their endogenously conceived, imagined experiences of the event. By asking questions that require introspective consideration and result in responses rich in information related to cognitive mechanisms concerning judgement (e.g., estimations) or decision making (e.g., hypothesising), the interviewer may be providing a liar with a framework with which to provide a plausible answer.

On a positive note, previous advantages of unanticipated questioning for observer judgements were confirmed in Experiment 2, particularly with unanticipated questions that focussed on spatial/temporal details. Moreover, the findings of the dichotomous decisions showed that, in line with the results of Experiment 1, the advantages of asking unanticipated questions was more evident for the detection of liars. This finding may have important real-world implications in certain legal situations. For example, in some word-against-word situations, as is often the case in sexual crimes, the ability of observers (e.g., jury members, judges, etc.) to accurate determine the veracity of an interviewee's statement may be vital in obtaining an accurate and successful outcome. However, the increase in cognitive load experienced by truth-tellers raises the concern that, if used in practical settings, insensitive use of unanticipated questioning may increase the likelihood of mistaking truth-tellers for liars. Spatial/temporal questions emphasise differences in the ways in which truth-tellers and liars use contextual words; planning questions that encourage the discussion of cognitive operations do not.

Taken together, the results of the studies provide little support for the unanticipated questions approach to veracity testing. There is some support for the approach in a post-interview observer scenario, though it appears that some forms of unanticipated question will be more successful in this situation than others. Furthermore, the cognitive load explanation provided by Vrij (2014) was refuted, leading to potential concerns regarding the application of the approach in practical settings.

Chapter 4: An exploratory analysis of interview question generation: The effects of training, topic, and temporal perspective

Abstract

Numerous investigative interviewing techniques offer empirically supported methods for asking investigatively useful questions, designed to elicit verbal cues to deceit. However, little research has examined the underlying dimensions of question quality. Moreover, there is no research examining the conditions in which individuals are able to generate the types of questions proposed by these techniques, nor the factors that enhance or inhibit that ability. This exploratory experiment sought to address these issues. Novice participants, half of whom had received training, were shown information-gathering interview clips in which interviewees discussed a number of general topics, either from a past or present temporal perspective. After each clip, participants were asked to generate an investigatively useful follow-up question. Subsequently, the questions were rated by experts for quality, as well as for novelty and utility (traditionally considered the main two dimensions of creativity; Finke, 1990). Results revealed that there was an interaction effect between the topic of interview and the temporal perspective of topic on the quality ratings. We argue that this is due to the scope of episodic information inherent within the topic/temporality combination. Additionally, training had a positive impact on question generation performance. However, the two dimensions of creativity proved to be an unreliable method of judging question quality. Further research is required in order to establish the underlying dimensions of question quality, and to fully explore the effect of scope.

Introduction

With the introduction of the PEACE model in 1993 came a realisation among law enforcement agencies that different types of question can have a crucial impact on interview outcomes. For example, research highlighted the utility of open questions at the beginning of interviews (Myklebust & Bjørklund, 2006) and warned about the complications that can arise when using leading questions (Oxburgh, Ost, & Cherryman, 2012). In the decades since, numerous interview techniques have been developed with a focus on the content of such questions, such as the Unanticipated Questions technique (UQ; Vrij et al., 2009), Strategic Use of Evidence (SUE; Granhag et al., 2007) and Controlled Cognitive Engagement (CCE, Ormerod & Dando, 2015). There is evidence to suggest that each of these techniques are effective tools for gathering information and distinguishing between true and false accounts (Hartwig et al., 2014; Ormerod & Dando, 2015; Sooniste et al., 2017). However, there is little to no research that has investigated the process of exactly *how* interviewers are able to generate the questions suggested by these methods. To this end, the present paper aimed to make an exploratory examination into the creative process of question generation, with the intention of developing a method for rating the quality of investigative interview questions.

As well as instilling ideas such as building rapport with the interviewee and outlining the objectives of the interview before starting (Authorised Professional Practice, 2019), one of the key improvements introduced by the PEACE model was to outline, and distinguish between, different types of questions and to show the impact those different question types can have on interview outcomes (Oxburgh et al., 2010). For example, the model encourages the use of open-ended questions at the beginning of interviews in order to allow the interviewee to provide a full, uninterrupted account that

is not influenced by the interviewer. This kind of free recall at the beginning of an interview has been shown to contribute a large proportion of all information obtained during an interview (Milne & Bull, 2003), can facilitate rapport (Read, Powell, Kebbell, & Milne, 2009), and can be used as a basis with which to structure subsequent lines of enquiry (Milne, Shaw, & Bull, 2007). Conversely the model warns about the use of leading questions, highlighting issues such as the potential to adversely influence an interviewee's response or distort their memory (Oxburgh et al., 2012).

Whilst this was a positive step forward for investigative interviewing, Chapter 3 showed that, even for questions that are of a similar type, the content of those questions can have a significant impact on the type of information received in response, which in turn can have an impact on the outcome of the interview. In the intervening decades since the introduction of the PEACE model, numerous techniques, such as the UQ approach (Vrij, et al., 2009), have been developed that place greater focus on the content of questions used in investigative interviews.

The UQ approach aims to raise the cognitive load of liars by asking questions that are not anticipated by the interviewees (Vrij, 2014). Given the opportunity to plan for an interview, liars will frequently attempt to anticipate what they are likely to be asked and prepare plausible sounding responses, often referred to as a 'lie script' (Colwell et al., 2007; Vrij et al., 2017). Vrij (2014) argues that, by designing questions that the interviewees have not considered, this removes a liar's ability to rely on their lie script and forces them to respond spontaneously, increasing their cognitive load. Several studies show that, in response to UQs, liars tend to provide shorter, less detailed and less consistent responses than truth-tellers (Sooniste et al., 2015; Vrij, Mann, et al., 2012). This suggests that the UQ approach can be a useful technique in veracity

detection and highlights the importance of focusing on the content of the questions asked.

Despite this empirical support for the UQ approach, there have also been studies that bring its efficacy into question, for example, finding no difference in the amount of detail provided by truth-tellers and liars in response to UQs (Vrij et al., 2011). Chapter 3's examination of the technique identified a number of potential issues. The experiment compared the use of anticipated questions to both UQs regarding the planning of an event and UQs focussed on the episodic details associated with the event. Results revealed that question type had no significant effect on the veracity detection accuracy of interviewers; UQs did not appear to improve interviewers' ability to distinguish between true and false accounts. Conversely, a follow-up study, in which participants were shown transcripts of the interviews and were asked to state whether they believed them to be truthful or deceptive accounts, did reveal a significant effect of question type. UQs improved observer veracity detection accuracy. However, the episodic detail UQs improved performance more than the planning UQs. Moreover, a Reality Monitoring analysis of the interviewees' verbal content revealed significant differences in the type of information gathered in response to the two UQ types. Overall, the results indicate that the content of UQs needs to be further considered and, moreover, they suggest that anticipation alone might not account for the reported improvement in veracity detection accuracy.

Another potentially useful interview technique is CCE (Ormerod & Dando, 2015), which was shown to yield high detection rates of mock airline passengers (70%+), in an aviation security field study where mock passengers were mixed in a ratio of 1:1000 with genuine passengers. The method incorporates UQs within its framework and was shown in a large-scale study conducted in a real-world aviation security setting

to be a consistently reliable method of veracity detection. So, what differentiates Vrij and colleagues' (2009) approach to UQs from Ormerod and Dando's approach? Firstly, CCE gathers together a number of empirically tested processes into one overarching technique, incorporating UQs within this paradigm. For example, CCE highlights the importance of initially building rapport, which has been shown to improve interview outcomes (Abbe & Brandon, 2013; Collins et al., 2002; Stokoe, 2009). It also stresses the importance of baselining; that is, beginning the interview with some straight-forward, easy to answer questions in order to establish how the interviewee behaves when not under pressure. This allows the interviewer to compare the interviewee's behaviour when relaxed to their behaviour when faced with more taxing UQs. Secondly, CCE places far more focus on the content of the UQs. The technique works by asking open-ended questions, allowing the interviewee to provide a free account, and then challenging them on aspects of that information. According to Ormerod and Dando, these challenges should be a 'test of expected knowledge', that is, something the interviewee should be capable of answering on the basis of their experience given the information they have just provided. Furthermore, this test of expected knowledge should be both unanticipated and focus on the episodic details associated with the topic in hand.

Both techniques mentioned above have been shown to be effective methods for investigative interviewing, and both, to varying degrees, place some focus on the content of the questions being asked. For example, we know from CCE that a good quality interview question should be an unexpected test of expected, episodic knowledge (Ormerod & Dando, 2015). However, what remains unaddressed is exactly *how* to generate a good quality test question such as this. As such, forensic psychology has over two decades worth of research describing how different types of questions

affect interview outcomes, but no research into the factors and conditions required to generate those questions in the first place.

The present study was designed as an exploratory first step in addressing this gap. Four individuals were interviewed about four topics (home, hobbies, work and travel). The topics were varied in temporality so that they were discussed from either a past or a present perspective. Subsequently, videos of the interviews were shown to a group of novice participants who were tasked with generating an investigatively relevant follow-up question after each interview clip; one which would explore the interviewee's presented account in a way that would best assist an investigator in assessing the veracity of the interviewee's account. Half of the participants were shown a 10-minute training video in advance, whilst the other half were not. Following this, each of the questions were rated for quality by two experts, in a preliminary effort to determine factors that may contribute towards good quality investigative questioning.

The CCE method of interviewing, as applied to aviation security, encourages the security agent to rotate through various topics when interviewing each passenger (Ormerod & Dando, 2015). There is no restriction on the topic of discussion, which can include education, hobbies, family, or any other topic the agent chooses. In theory, this prevents a deceptive interviewee from predicting what information to prepare in advance. As such, the first variable explored in the present study was interview topic, in an attempt to determine whether certain topics elicit questions rated as higher quality (in terms of the value of the answers they might yield to investigators/security screeners) than others. The technique also promotes the rotation of temporality, encouraging the agent to vary the time-line of discussion, for example, asking for information regarding the passenger's past education, followed by asking about their current job. Doing so increases the amount of information a deceptive interviewee is required to generate and

deliberate on, since they must create and maintain the consistency of both past, present and future falsehoods. Therefore, temporality was also manipulated in order to investigate its effect on question generation, in isolation as well as in its interaction with topic.

Finally, training was manipulated. There is evidence to suggest that short video-based training can improve performance on technical tasks (Maldarelli et al., 2009; Truebano & Munn, 2015). However, this has not previously been applied to question generation. By presenting half of the participants with a brief training video, the present study explored whether the skills described by the CCE technique are capable of being conveyed in such a manner, and whether they improve performance in a question generating task.

Generating questions such as the test questions suggested by the CCE technique is an inherently creative task. It requires the interviewer to listen carefully to the interviewee's account, pick out key details, consider the range of episodic information that one could reasonably expect the interviewee to know based on their account, choose an aspect of that episodic information that the interviewee would not necessarily anticipate being asked about, before finally using that information to formulate a test question. Therefore, to investigate this ability, it is logical to also explore it from a creativity perspective. Research into creativity generally relies on a standard definition that incorporates both novelty and utility (Finke, 1990; Sternberg & Lubart, 1999). For example, if looking at designs for a new product, in order to be considered creative the product should be in some way unique. However, originality alone could conceivably be achieved purely by random generation of worthless artefacts. Therefore, to be considered truly creative, the product must be capable of effectively achieving its intended purpose (Runco & Jaeger, 2012). Arguably, this definition maps onto

investigative question generation. Taking the CCE approach, a good quality test question will be one that the average interviewee would not have considered themselves (i.e., it is novel), and will test their episodic knowledge in a way that helps to establish the veracity of their account (i.e., it is useful).

Therefore, in addition to assessing global ratings of the questions' general quality, the present paper sought to make use of the standard definition of creativity in order to explore the creative process of question generation. Finke (1990) pioneered the use of novelty and utility ratings in measuring the creative value of imagined objects. For example, Verstijnen, van Leeuwen, Goldschmidt, Hamel, and Hennessey (1998) presented participants with a series of basic shapes (e.g., cube, sphere, cylinder, etc) and asked them to use the shapes to design and name a creative object. The objects were subsequently rated by five judges for originality and practicality. Runco and Charles (1993) conducted a study to assess the contribution of originality and appropriateness on overall creativity ratings in divergent thinking tasks. The results showed that the subjective ratings mapped onto objective measures of creativity. They argued that originality contributed more towards creativity than appropriateness, though when levels of both increased, subjective ratings of creativity did also. However, McKnight, Ormerod, Sas, and Dix (2006) found that originality and practicality were often in opposition: when given time to explore objects before commencing designing, participants produced designs that were more original but less practical, and vice versa when they began to design immediately. Given that this procedure has never been applied to question generation, five experts were asked to make novelty and utility ratings for the present study, in line with Verstijnen et al (1998).

Research into creativity exhibits a great deal of variation in terms of rating methods employed and inter-rater agreement between expert judges, often due to the

subjective nature of judging creativity (see Long, 2014). Amabile's (1982) Consensual Assessment Technique (CAT) of creativity rating suggests that judges often make more reliable ratings when making global judgements of creativity, as opposed to rating explicitly defined facets of creativity. As such high levels of inter-rater reliability were expected to be found between the two experts' ratings of general quality (Hypothesis 1). The participants that had been trained before doing the task were expected to generate questions that were rated higher in quality, than those that had not been trained in advance (Hypothesis 2). In terms of the creativity ratings, in accordance with Finke (1990), the five judges were provided with clear, objective definitions of novelty and utility in relation to the CCE model of test questions. Therefore, high levels of inter-rater reliability were expected to be found between the five experts' ratings (Hypothesis 3). Given that the CCE method suggests that a good quality test question will be both unanticipated and a useful test of expected knowledge (Ormerod & Dando, 2015), there should be a positive correlation between the quality ratings and the novelty ratings, as well between the quality ratings and the utility ratings (Hypothesis 4). At this stage we did not make any firm predictions about the effect that interview topic and temporal perspective of the topic might have on creativity and quality ratings.

In the study described below, we first generated a set of short interview videos in which actor interviewees discussed four topics: home, hobbies, work and travel. The topics were discussed either from a past or present temporal perspective. Videos of these interviews were subsequently used as the stimulus for a group of novice participants to generate follow-up question which would challenge the interviewees' accounts. These questions were presented to two expert interviewers who rated each for general quality. Additionally, five interviewing experts rated the questions for novelty

and utility. Overall, the present study sought to offer an initial exploration into the creative process of investigate question generation.

Method

Participants

Interviewees. Two males and two females were selected to serve as actors and be interviewed about their hobbies, home, work and travel plans. Interviewee HT was aged 33 and was a PhD researcher; CG was aged 46 and was a part-time PhD researcher and part-time school governor; CK was aged 24 and employed as a research assistant; and TT was aged 31 and employed as a lift engineer. All agreed to be interviewed voluntarily, without financial incentive.

Question Generators. Forty-four females (M_{age} = 21.18 years, SD = 2.94) and sixteen males (M_{age} = 22.25 years, SD = 6.81) were asked to generate follow-up questions for the study. Participants were UG and PG students from a range of science and arts disciplines at the University of Sussex. Each received either £5 or course credits for taking part in the study.

Question Raters. Two male interviewing experts were recruited to rate the general quality of the questions. Additionally, two further female and one male interview experts were recruited to rate the creativity. Three of these were from an academic background: Rater 1 (quality and creativity ratings) was aged 56 and had 6 years of interviewing experience at Tier 1 level; Rater 2 (quality and creativity ratings) was aged 33 and had 4 years interviewing experience. Rater 3 (creativity ratings) was aged 32 and had 6 years of interviewing experience having completed Achieving Best Evidence training. The remaining two were from a law enforcement background: Rater 4 (creativity ratings) was aged 51 and had 25 years of experience as a Tier 5 home

office interview advisor; Rater 5 (creativity ratings) was aged 57 and had 35 years of experience as a criminal justice consultant.

Design

A mixed design was employed for the interviews. All four interviewees were asked to speak about each of the four topics; hobbies (n = 4), home (n = 4), work (n = 4) and travel (n = 4).Temporality was systematically varied so that they were asked about each topic from perspective of either the past (n = 8) or the present (n = 8). This produced 16 interview videos in total.

A mixed design was also employed for the question generating element of the study. Participants were each shown all 16 of the interview videos and were asked to generate a question after each video. Training was systematically varied so that half of the participants watched a training video before taking part in the task (n = 30) and half were offered the opportunity to watch the video after taking part (n = 30).

Procedure

The interviews. Each of the interviewees was asked to speak about four topics from the perspective of either the past or the present. For the present perspective conditions, the interviewer framed the questions as follows: “what hobbies or leisure activities do you currently take part in”, “tell me about where you live at the moment”; “tell me about any travel plans you have coming up”; and “tell me about the job you’re currently doing.” The interviewer then allowed them to respond to the question and, if necessary, would prompt them for more information simply by saying “please tell me more about that.” All participants were asked to respond truthfully to all of the questions. For the past perspective conditions, the interviewer framed the questions as follows: “tell me about a hobby you used to do, but don’t do anymore”; “Describe

where you were living 10 years ago"; "Tell me about the first overseas holiday you remember going on"; and "Tell me about the first job you had." Again, the interviewer allowed them to answer the questions, prompting them for more information if necessary, and the interviewees were asked to answer honestly.

Each of the interviews was filmed. Subsequently, videos of the interviews were edited so that each contained the question and response of one interviewee to one topic, creating 16 videos in total. The videos were then further edited to remove superfluous information, such as off-topic conversation between the interviewer and interviewee, and to attempt to make sure that each video was of a similar length and contained a roughly similar amount of information points.

Question Generation. Participants were informed that they would be watching 16 interview videos in which an interviewee would discuss one of four topics. They were instructed to imagine that they were the interviewer and that the goal of those interviews was to establish whether the account given by the interviewee in each clip was a true or false account. To this end they were asked to generate a question after each clip that they felt was creative, would provide some useful information and would challenge the interviewee's account.

Half of the participants (referred to hereafter as 'Trained') were shown a 10-minute training video designed to assist them in generating good quality, creative questions. The training video was created by the authors and used layperson's terms to cover some of the fundamental aspects of the CCE interview technique (Ormerod & Dando, 2015). The video covered elements of the technique such as the use of unanticipated questions, tests of expected knowledge and the distinction between semantic and episodic knowledge. The other half of the participants (referred to

hereafter as ‘Untrained’) were not shown this video before taking part but were offered the opportunity to watch it upon completion of the study.

The experiment was presented to participants on a computer screen using Qualtrics software. After being presented with the study information, consent form, instructions and the training video (in the case of the Trained participants), they were asked to provide basic demographic information and then the main task began. The videos were presented to them on screen in a random order. After each video, the screen automatically proceeded to the question generation screen, with a box for them to type the question they had generated for that video. They were given 60 seconds to generate a question, at which point the screen automatically moved on to the next video. This procedure was repeated until all 16 videos had been viewed.

This process generated a total of 960 questions. However, there were a number of duplicate questions. The lead author categorised the duplicate questions for each of the 16 videos and then randomly chose one of each set of duplicates to be included in the rating section of the study. The remaining duplicates in each set were allocated the same rating as the one chosen to be rated. In total, 596 of the 960 questions were included in the creativity rating section of the experiment.

Question Rating. The expert raters completed the task online using Qualtrics software. Initially, they were asked to fill out a demographic questionnaire that asked them their age, profession, years of interviewing experience, and level of interviewing achieved. They were then presented with the instructions for the task. They were told about the question generation task and were then informed that they would be shown the 16 interview clips followed by the questions that had been generated for each video.

Two experts rated each of the questions in terms of the 'general quality' of the questions on a 1-7 scale (1 = very poor quality; 7 = very good quality). For each set of questions, the rater was first presented with the interview video on screen. They were also provided with a Youtube link to the video so that they were able to re-watch it if they wished. Following the video, they were presented with each of the questions generated for that interview, in a random order. Questions were presented one at a time, with the scale for rating quality presented below the question. When they had completed one set of questions, they were randomly provided with another interview video and its associated questions. This was repeated until all 596 questions had been rated.

Additionally, five experts were asked to rate each question on a 1-7 scale for both novelty (i.e., the level to which they felt the interviewee might have expected to be asked the question) and utility (i.e., the ability of the question to determine the veracity of the interviewee's account). This was performed using the same method presented above for the general quality ratings.

Intraclass Correlation Coefficient Analysis

Inter-rater reliability for the novelty, utility and general quality ratings was calculated using Intraclass Correlation Coefficients (ICC). In each instance a two-way random-effect model based on average ratings and consistency (ICC(C,k)) was applied, as the intention was to calculate the mean of the judges' ratings for each scale. Advice on interpreting ICCs is mixed. Cicchetti (1994) indicates that a coefficient between .60 and .74 can be interpreted as 'good' and above .75 as 'excellent'. Koo and Li (2016) offer different guidelines, suggesting that a coefficient between .50 and .74 is just 'moderate' and above .75 is 'good'. For the present paper, we followed Cicchetti and Sparrow's (1990) suggestion that scores below .70 demonstrate an unacceptable level of

agreement and, therefore, chose only to include scales that met this threshold in subsequent analysis.

Results

General Quality Ratings

Reliability. In order to determine the level of inter-rater reliability between the two raters, an ICC(C,k) analysis was conducted. The ICC was .80, 95% CI [.77, .83], which indicates good reliability and meets the acceptable threshold. Therefore, the two sets of ratings were merged into one set of mean ratings for question quality. All subsequent analyses were performed using this mean quality rating.

Training. The 16 questions generated by each participant were rated for quality on a seven-point scale (1 = very poor question; 7 = very good question). An overall mean quality score for each participant was calculated. Figure 1 shows the mean quality rating for questions generated by participants in Trained and Untrained conditions.

In order to determine whether the training video had an effect on the quality of the questions generated by the participants, an independent t-test was carried out on the mean level of question quality. The results showed that the participants who had watched the training video before taking part in the task provided significantly better quality questions (M = 3.32, SD = 0.76) than those who did not watch the training video before taking part (M = 2.62, SD = 0.65), $t(58) = 3.86$, $p < .001$, $d = 0.99$, 95% CI [0.45, 1.52].

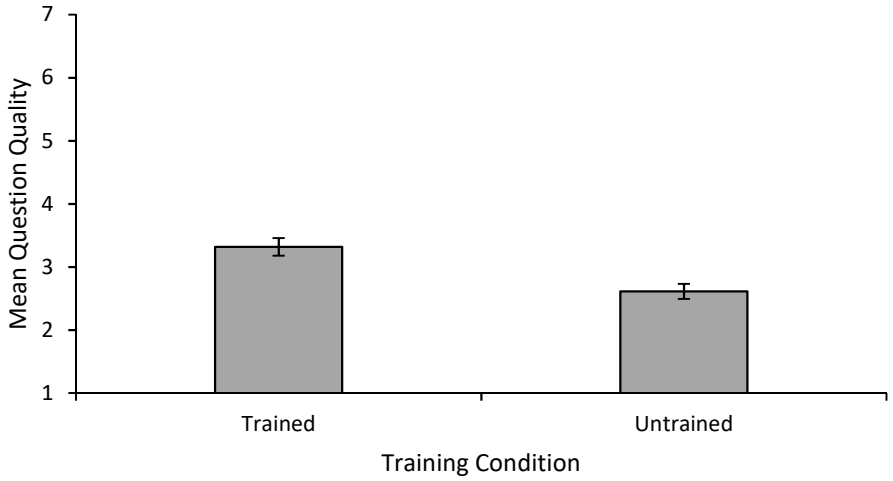

Figure 1. Mean rating of quality for questions generated by trained and untrained participants. Error bars represent +/- 1 SEM.

Topic and Temporality

In order to determine whether the topic of the interviews and the temporality in which the topics were framed had an effect of the quality of questions generated, a 4 (topic: hobbies, home, travel, work) × 2 (temporality: past or present) repeated-measures ANOVA was conducted. There was a significant main effect of topic, $F(3, 177) = 10.15$, $p < .001$, $\eta_p^2 = .15$, 95% CI [.03, .20]. There was not a significant main effect of temporality, $F(1, 59) = 2.92$, $p = .09$. However, there was a significant interaction between topic and temporality, $F(3, 177) = 6.44$, $p < .001$, $\eta_p^2 = .10$, 95% CI [.03, .19].

In order to examine this interaction further, a series of follow-up repeated-measures t-tests were conducted (see Figure. 2). There was no significant difference in quality ratings found between the questions generated for the interviews concerning hobbies in the past (M = 2.78, SD = 1.05) and the present (M = 2.63, SD = 1.01), $t(59) = 1.24$, $p = .22$. Nor was there a significant difference in quality ratings found between

the questions generated for the interviews concerning home in the past (M = 2.97, SD = 0.92) and the present (M = 2.92, SD = 1.35), $t(59) = 0.26$, $p = .80$. However, for the questions generated regarding travel, quality was rated as significantly greater when the topic concerned the past (M = 3.58, SD = 1.48) than the present (M = 3.01, SD = 0.87), $t(59) = 3.78$, $p < .001$. The converse was true for the questions generated regarding work, with those framed in terms of the past (M = 2.92, SD = 1.08) being rating as significantly lower in quality than those concerning the present (M = 3.16, SD = 0.87), $t(59) = -2.08$, $p = .04$.

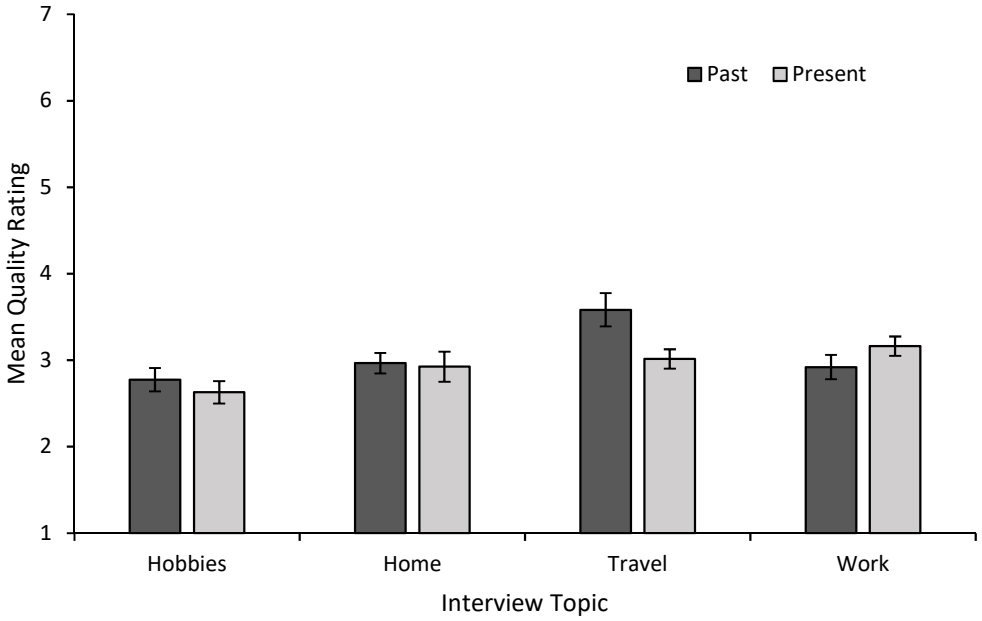

Figure 2. Mean quality rating of questions generated in each topic, framed both in the past and the present. Error bars represent +/- 1 SEM.

Creativity Ratings

Reliability. In order to determine whether it was appropriate to merge the five experts' ratings into single measures of novelty and utility, three ICC(C,k) analyses were conducted. For the 596 novelty ratings made by the five raters, the ICC was .65, 95% CI [.60, .69]. This indicated moderate inter-rater reliability but failed to meet the

threshold for acceptable reliability. For the 596 utility ratings made by the five raters, the ICC was 0.61, 95% CI [.56, .66]. This indicated moderate reliability and did not meet the threshold. The novelty and utility scores were summed to make an overall rating of creativity. For this variable, the ICC was .65, 95% CI [.61, .69]. Again, this indicated moderate reliability and did not meet the threshold.

In order to determine whether this poor reliability was due to differences between the ratings made by the practitioner and academic expert raters, a further set of four ICC analyses were conducted using the same model outlined above. For the two practitioner raters, the ICC for novelty and utility were .31, 95% CI [.19, .41] and .34, 95% CI [.23, .44], respectively. For the three academic raters, the ICC for novelty and utility were .66, 95% CI [.61, .70] and .51, 95% CI [.43, .57], respectively. Given the poor-to-moderate inter-rater reliability indicated between both sets of raters, it appears that the raters' profession was not able to account for the below-threshold inter-rater reliability found overall. Therefore, we were forced to conclude that the expert raters were in disagreement with regards to the ratings of both novelty and utility. This meant that the ratings could not be merged to perform any subsequent analysis.

Relationship between creativity and quality. Whilst it was not possible to conduct further analysis with the creativity ratings, as two of the raters had made judgments using both the general quality scale and the creativity scales, it was possible to analyse the relationship between those scales within their individual scores. As can be seen in Table 1, for both raters there was a strong correlation (Cohen, 1988) between their quality ratings and overall creativity ratings. However, for both raters, the correlation is almost entirely as a result of the utility ratings. Correlations between the raters' quality ratings and their novelty ratings were weak in both instances, implying that the raters did not consider novelty to be a useful indicator of quality.

Table 1.

Correlation between each rater's general quality rating and their own rating of novelty, utility and overall creativity.

	Novelty	Utility	Overall Creativity[1]
Rater 1	.17	.71	.61
Rater 2	.24	.75	.66

Note: [1] Overall creativity calculated as sum of novelty and utility.

Discussion

The experiment examined the process of question generation in novice participants. An exploratory approach was undertaken, which investigated the effect of interview topic, temporal perspective of topic and training on the quality and creativity of investigate question generation. The findings confirmed the initial hypothesis, in that the experts' ratings of general quality showed good inter-rater agreement. Using the general quality ratings, the findings indicated that participants who had watched a short training video in the CCE technique generated higher quality questions than those that were untrained. This supported our hypothesis, as well as providing support for CCE technique more broadly. There is evidence to suggest that short video-based tutorials can have a positive effect on performance in technical tasks (Maldarelli et al., 2009; Truebano & Munn, 2015), and the results increase our confidence in the training video as a tool for future question generation studies. Given that the training video described methods associated with CCE (Ormerod & Dando, 2015), with a particular focus on generating unexpected tests of expected episodic knowledge, these findings can be taken as support for the efficacy of the technique in terms of improving an interviewer's ability to generate good quality questions. It also demonstrates that the principles of the technique can be conveyed succinctly and effectively, even to novices with no experience of investigative interviewing.

The exploratory analysis into topic and temporality revealed that for interviews regarding travel, framing the questions in a past perspective results in higher quality questions than the present, whilst for the work interviews the opposite was true. No significant difference in temporal perspective was found for either the home or the hobbies interviews. Given the exploratory nature of this study, it is only possible to speculate on the reason for these findings at the present time. However, one could argue that they make intuitive sense if viewed in terms of the breadth of episodic information inherently associated with each.

Considering first the travel interviews, the past perspective question required the interviewee to discuss the 'first overseas holiday they remember going on'. For most people, their first overseas holiday is an important, memorable event. Even if it occurred a long time ago, they are likely to remember a significant amount of information concerning that trip and a large proportion of that information will be episodic. In contrast, the present perspective question required the interviewee to discuss 'any travel plans they have coming up'. In this instance, the interviewee has yet to develop an episodic memory of the event, given that they have not yet experienced the trip. Therefore, the only episodic information available for the interviewer to tap into is based on the planning of the trip. In chapter 3, reality monitoring analysis showed that asking planning-based questions results in responses that are rich in 'cognitive mechanism' words (i.e., words associated with thought processes), and this in turn can have a negative impact on interview outcomes.

For the work interviews, the past perspective questions asked the interviewee to talk about the 'first job they had'. For a lot of people, and indeed the interviewees in the present study, their first job tends to be at a low-level with little responsibility. Most people are likely to retain some episodic memory of their first job. However, when

compared to their current employment, the breadth of episodic information available is naturally going to be limited. The finding that temporal perspective had no effect in the home interviews arguably fits into this theory. Given the amount of time we spend in our homes, it is reasonable to assume that a similar level of episodic information is available whether we are discussing our current home or the one we lived in 10 years previously. Temporal perspective also had no effect in the hobbies interviews, which is potentially incongruous with the theory. Intuitively, one might expect to find a richer depth of episodic information available when discussing hobbies that the interviewee currently engages in, compared to hobbies they used to engage in. However, when asking someone to describe a hobby 'they used to do, but don't anymore', they are likely to consider whichever hobby was most influential in their past. Whereas, they may have only recently begun to take part in their current hobby. Therefore, in terms of the amount of available episodic information to draw on, the recency effect of current hobbies might be cancelled out by the depth of experience associated with past hobbies.

The exploratory investigation into the effects of topic and temporality on participants' ability to generate good quality interview questions suggests that the scope of episodic knowledge inherently available to the interviewer, based on the topic of discussion, might be a crucial component. When the topic of investigation provides a wide scope of episodic information with which to formulate questions, the quality of questions should be greater than when the topic of discussion provides only a narrow scope. However, it must be noted that this theory is entirely speculative at this stage and will require empirical testing in subsequent studies. This could be achieved using a similar method to the present study but with a controlled manipulation of the scope of the interview topics. For example, participants could be shown a series of interviews, half of which have been conducted in a way that limits the scope of available episodic

information and half of which that have been conducted in a way to increase the scope. Participants would generate follow up questions after each that would then be rated to determine whether the broad scope interviews elicit higher quality test questions than the narrow scope interviews.

There was a strong correlation between the two judges' ratings of general quality and their ratings of creativity, however, this effect was largely due to the utility ratings as opposed to the novelty ratings. Novelty appears to be less relevant than utility when judging the quality of investigative questions. Despite these findings, the five expert judges were low in agreement for both novelty and utility, therefore, any further analysis using these creativity scores was abandoned. Despite being provided with clear, objective definitions of novelty and utility, the expert judges did not agree on either. Moreover, this disagreement was not explained by the judges' basis of expertise, with equally poor agreement found between the ratings of the two practitioner experts and the three academic experts. A potential reason for this may be found in the issues with regards to defining expertise. Each of the expert raters had at least 5 years' experience in a relevant field, but experience does not necessarily equate to expertise. However, it is perhaps a more likely explanation that opinions on the creative value of test questions remain subjective and open to the individual expert's interpretation of novelty and utility, even when these concepts have been plainly defined. There is precedent for this in the creativity literature, with inter-rater reliability being the source of much debate in the field.

Hickey (2001) had five groups of judges rate the creativity of children's musical compositions. The groups of judges ranged in expertise from professional composers to second grade children. The results showed that the composers were the least in agreement of any of the groups. Similarly, Runco, Mccarthy, and Svenson (1994) had

professional artists and novices rate the creativity of three pieces of art. Whilst the expert judges showed a reasonable level of agreement, their ratings were not able to distinguish between the three art pieces (i.e., there was no significant difference in ratings between the three works). Furthermore, their ratings were significantly lower than the novices' ratings. The authors concluded that the experts were overly critical and were insensitive to the differences between the pieces they were rating. Amabile (1982), who developed CAT, argued that by attempting to define creative constructs objectively, when conducting rating studies such as these, researchers "often fail to differentiate between the creativity of the products and other constructs, such as technical correctness or aesthetic appeal. Moreover, the interjudge reliability might be questioned in studies where the experimenter presents judges with his or her own definitions of creativity to apply" (p. 1000).

In response to these concerns, many creativity theorists argue that it is more prudent to provide judges with subjective, implicitly defined measures of creativity with which to make their ratings. Amabile (1982) developed CAT on this basis, arguing that whilst there are certain characteristics that undeniably contribute towards creativity, such as novelty and utility, the choice and definition of these characteristics is ultimately subjective in nature. Amabile supported this theory in a series of experiments which showed that high inter-rater reliability was observed when experts were provided with implicit definitions of creativity and were given licence to apply their judgements subjectively. Further support was provided by Webster and Hickey (1995) who showed that, when judging the global characteristics of children's musical compositions, ratings were significantly more reliable when the constructs were implicitly defined than when specific, explicit definitions were used. This form of rating is more in line with the principle form of ratings used in this study, whereby the experts rated each question

simply in terms of general quality. In principle, this method was far more subjective and less clearly defined. However, the results showed that the inter-rater reliability was moderate to good, supporting the first hypothesis. This suggests that the CAT approach may be a more appropriate method for rating question quality.

The experiment used novice participants to generate the test questions. Whilst the CCE training video did have a positive effect on their performance, the mean quality rating for the trained group was still below the mid-point of the scale, that is, they still generated questions that, on average, were considered to be relatively low in quality. In order to gain a more complete picture of the factors involved in question generation, it will be necessary to explore ability among individuals with existing experience of investigative interviewing. It would be reasonable to expect participants with this experience to generate higher quality questions, which in turn would allow for further investigation into the factors that contribute towards high quality questions.

Another potential limitation of the study was the use of the training video, and the conclusions drawn from its use. The results indicated that those who watched the video generated significantly higher rated questions than those participants who had not watched the training video. We took this as support for the efficacy of the training video as a tool for improving the performance of novice participants, and also as an indication that the CCE method of interviewing (Ormerod & Dando, 2015) is able to be conveyed in a clear, succinct manner. However, there is a possible that the increased level of performance may have simply been due to a motivation effect. As such, future studies looking into the effects of training videos should perhaps use a control video (e.g., covering a different technique), in order to determine whether it is the material in the training video that had an effect on the outcome, or simply watching a training video in general.

The exploratory investigation into question generation revealed that explicitly defined concepts of novelty and utility were not an adequate measure by which to judge question quality. The results showed that expert judges were not able to agree on these measures, suggesting that these characteristics are subjective in nature. Whilst the more subjective, global ratings of quality did prove to be a more reliable measure, applying such a subjective measure in a practical setting is limited in terms of the ability to explore the components that contribute towards good quality questioning. As such, developing a novel approach to question rating, which takes into account components that are more directly relevant to interview question quality, might be necessary. This will require further exploratory examination in order to determine such components. Any technique would need to be compared to the subjective ratings of quality, given that currently this is our most reliable method.

Taken together, the findings of this investigation suggest that much more exploration is required in order to build a more complete picture of the creative process of question generation. Literature in the creativity field applies a standard definition of creativity that incorporates novelty and utility (Finke, 1990; Sternberg & Lubart, 1999). This proved to be an unreliable measure by which to measure question quality in the present study. Applying the more implicitly defined measure of general quality was an improvement. However, this does not reveal the components that contribute towards question quality. Therefore, novel techniques need to be explored. The findings showed that there was an interaction between topic of interview and temporal perspective of topic on subsequent question generation quality. We proffer a speculative theory that the scope of available episodic information inherently available to the interviewer offers an explanation for this interaction. Further research will seek to test this theory

empirically, as well as investigate novel methods for assessing the quality of investigative interview questions.

Chapter 5: Applying a bottom-up approach in search of the underlying dimensions of question quality

Abstract

There is a wealth of research exploring techniques for improving the quality of investigative interviewing, but little attempt has been made to establish the dimensions that separate good- and poor-quality questions. As a result, there is no method for objectively measuring the quality of investigative questions. This two-part experiment was devised in order to develop such a rating method, as well as to explore certain factors which may affect the ability to generate good-quality questions. In a pilot study, a card sort was applied to a large sample of novice-generated questions. This revealed a 3-dimensional model of question quality comprising Relevance, Unpredictability and type of Knowledge probed. In the main study, novice and expert participants were shown a number of interview clips and were required to generate an investigatively useful question after each. The incidents discussed in the interviews varied in terms scope and the veracity of the interviewee. Novice and expert judges rated the questions for subjective general quality. Additionally, the experts made ratings using the 3-dimensional model. The scope of episodic information was shown to be a context-dependent factor affecting the ability to generate good-quality questions. Expertise of the question generator and interviewee veracity had an effect on novices' ratings only, suggesting that expert and novice judges value different attributes when judging question quality. The 3-dimensional model is a potentially useful tool for judging interview questions, though further research is required in order to fully investigate its efficacy across various interview contexts.

Introduction

In the past two decades, a wealth of research has explored investigative interviewing techniques, resulting in a number of empirically backed methods designed to improve interview outcomes. For example, the Unanticipated Questions technique (UQ; Vrij et al., 2009) suggests the use of questions that interviewees are unable to prepare for in advance, whilst the Controlled Cognitive Engagement method (CCE; Ormerod & Dando, 2015) proposes using unexpected tests of expected knowledge. Whilst there is support for the efficacy of these questioning techniques, no research has focussed on the conditions required to generate such questions. In chapter 4, we explored the use of subjective general quality ratings, as well as creativity ratings, to assess the quality of novice-generated questions. The results showed that the topic of interview and the temporal perspective of the topic can influence the quality of subsequent test questions generated by participants. We theorised that the factor responsible for this influence is the scope of episodic information inherently available to the interviewer. The present study sought to test this theory, by manipulating the scope of interview discussions and then asking novice and expert participants to view the interviews and generate follow-up test questions. Furthermore, we explored a new method for rating the quality of investigative questions, as well investigating the effect of interviewee veracity on subsequent question quality.

Ormerod and Dando's (2015) 18-month field study, applying CCE to aviation security in a 1:1000 mock:genuine passenger context, showed that agents using the CCE technique identified 24 times as many mock passengers as agents using the currently employed method of security screening, which involves detecting ‘suspicious signs’ of behaviour (Martonosi & Barnett, 2006; Reddick, 2004). Bond and DePaulo's (2006) meta-analysis into veracity detection showed that humans’ ability to detect deceit rarely

exceeds chance. Even among trained investigative professionals rates tend not to be improved (Aamondt & Custer, 2006). As such, the findings of Ormerod and Dando (2015) can be taken as a positive step towards improving interviewers' ability to conduct good quality, effective investigations. Such advances are crucial given that the ability to distinguish between a true and false account can have important ramifications in law enforcement, such as gaining reliable information concerning potential terrorist attacks, preventing guilty suspects from avoiding charges, or acquiring accurate information from witnesses and victims (King & Dunn, 2010).

Despite the advances made by the development of methods such as CCE, as well as other potentially effective techniques like the UQ technique (Vrij et al., 2009) and the Strategic Use of Evidence approach (Granhag et al., 2007), no published research to date has investigated the conditions involved in, and required for, generating the types of questions put forward by these techniques. In Chapter 4, we made an initial exploration into question generation. Novice participants were shown a series of short information-gathering interview clips in which the interviewees discussed four topics (home, hobbies, travel and work), either from a past or present temporal perspective. After each clip the participants were asked to generate a follow-up test question that was relevant, useful and creative. Two interviewing experts rated the questions using a subjective measure of general quality. Additionally, five interviewing experts rated the questions generated by the participants for both novelty and utility- the two components considered to contribute towards the standard definition of creativity (Finke, 1990).

The two experts' ratings of general quality showed good inter-rater reliability and as these two experts had applied both rating schemes to the questions, it was possible to explore the correlations between their individual ratings. The findings revealed that there was a strong correlation between their ratings of utility and general

quality, but no such correlation when looking at novelty. This implies that the utility of a question is markedly more important in an investigative context than its novelty. Whilst novelty and unexpectedness are not necessarily directly synonymous, the raters were asked to consider the extent to which the interviewee might have expected to be asked each question when making their ratings of novelty, which brings into question the efficacy of techniques such as the UQ approach (Vrij et al., 2009). Furthermore, the study reported in Chapter 3 suggests that anticipation alone is not able to account for any improvement in veracity detection ability, and the findings of Chapter 4 appear consistent with this finding.

Using the general quality ratings in Chapter 4, the results indicated that there was a significant interaction between the topic of the interview and the temporal perspective of that topic. There was no difference in quality found between questions generated for past and present perspective when the topic was either home or hobbies. However, for travel interviews, questions generated for the interviews that were focused on the past were rated as higher in quality than those generated for interviews focused on the present. Conversely, for work interviews, a present perspective resulted in higher quality questions being generated. In the discussion of Chapter 4, we proffered the theory that these differences were potentially due to the scope of available episodic knowledge inherent in the context of the interview. Discussing travel in the present temporal perspective limits episodic enquiry to the planning of that travel, given that the event itself is yet to occur. Whereas, if discussing travel in the past, there is a rich vein of episodic knowledge to tap into. Regarding work, it is logical to assume that an individual will have more episodic knowledge to draw on when discussing their current job than a job they held many years ago.

This theory is arguably supported by Baddeley's (2012) model of working memory. Sporer (2016) applied the model to deception research, highlighting the importance of episodic knowledge in distinguishing between true and false accounts, noting that when an individual is preparing to tell a lie, if they do not have a true experience stored in episodic memory that is similar to the event they plan to lie about, then they have no option but to rely upon their semantic memory. As such, true accounts are likely to be far richer in episodic detail than imagined accounts, which should become evident to an observant interviewer.

Graesser (1981) argues that any event we experience is stored as part of a schema. Within that, there are schema-consistent details (i.e., details we normally associate with that event), schema-inconsistent details (i.e., details that are unusual for that event), and schema-irrelevant details (i.e., details that are considered not relevant to event). According to Sporer (2016), the longer a memory of an event is stored, schema-consistent details become more generic, whilst schema-irrelevant details tend to fade. However, schema-inconsistent details, given their uniqueness, tend to create a stronger memory trace and are therefore more likely to be recalled. This has an important implication for investigative interviewing as it suggests that lies are likely to consist of schema-consistent information and lack the unique, schema-inconsistent details. This maps on to the findings of Chapter 4 in the sense that, if the scope of the interview is such that it allows the interviewer to generate questions pertaining to this schema-inconsistent information, such questions are likely to be considered better quality than if the interviewer is limited to discussing schema-consistent information.

The present study was developed to test this theory. Four related tasks concerning the copying and distribution of a mock exam paper were designed, in which the scope of available episodic information was varied systematically. In narrow scope

versions of the tasks, participants were provided with specific instructions regarding the completion of the task in order to restrict the participant from exploring other available solutions. In the broad scope versions of the tasks, participants were encouraged to carry out the task however they wished, widening the scope of potential episodic information available to draw on. Participants completed each of the four tasks; two using the narrow scope instructions, and two using the broad scope instructions. Veracity was also manipulated so that they completed two of the tasks truthfully and two of the tasks deceptively. Moreover, a distinction was made between two forms of deception: lying to hide an act that has taken place and lying to pretend an act has taken place that has not. Memon and colleagues' (2013) unpublished literature review demonstrated that this difference can have an effect on the verbal behaviour of interviewees and thus affect interview outcomes.

The participants were subsequently interviewed regarding their involvement in the tasks. Using the videos generated from these interviews, two experiments were devised. A pilot study was undertaken in order to establish a novel method for rating question quality. Subsequently, the main study required novice and expert participants to view the interview videos and generate follow-up test questions. The questions were rated by experts using the novel rating system developed in the pilot study. Additionally, the experts and a group of novices made subjective general quality ratings. The study had three main aims: the initial aim was to develop a novel, objective and reliable measure of question quality that could potentially have a practical use. The second aim was to test the theory that the scope of episodic information available to the interviewer influences the quality of questioning. The final aim was to explore the potential effect that experience (of the question generating participants) and interviewee veracity has on question quality.

Pilot Study

Chapter 4 revealed that the implicitly defined measure of quality was a more reliable rating method than the explicitly defined measures of creativity. The two creativity scales, novelty and utility, exhibited poor inter-rater agreement between the five expert judges. There is evidence to suggest that explicitly defined scales are not necessarily appropriate for measuring creativity. Amabile (1982) argues that, by defining measures of creativity objectively, the researcher imposes their own definition of creativity on the judges and risks overlooking unspecified, but equally valid, creative components. The finding in Chapter 4 that the subjective measure of general quality resulted in good inter-rater agreement was more consistent with Amabile's Consensual Assessment Technique that requires judges to rate creativity in a more global sense, applying their own definitions to the ratings.

A difference between rating the creativity of objects, fictional writing, musical compositions, etc., and rating the quality of investigative interview questions is that there are an important, practical implications for question quality. As outlined in the general introduction, an inability to ask effective questions in investigative interviews can have severe implications (King & Dunn, 2010). As such, there is a need for an objective, easily-applicable measure of question quality in order to provide useful, practical advice to the investigative community. In turn, having an objective measure would allow for further exploration into the components and processes that contribute towards good quality, effective questioning.

To this end, a pilot study was designed in order to explore objective elements of good- and poor-quality questions. A group of novice participants were shown a series of short information-gathering interview clips and were required to generate a follow-up

test question after each. Following this, the authors conducted a series of card sorts, using the generated questions, with the aim of identifying a number of distinct question types. It was hypothesised that it would be possible to ascertain dimensions common to these distinct types that might distinguish between good- and poor-quality questions.

Method

Participants

Interviewees. Two males and two females were selected to carry out the tasks and subsequently be interviewed. Interviewee HT was aged 34; JM was aged 25; CG was aged 47 and AW was aged 24. All were PhD researchers at the University of Sussex and agreed to take part voluntarily, without financial incentive.

Question generators. Forty-nine female (M_{age} = 20.61, SD = 2.87) and 11 male (M_{age} = 25.55, SD = 8.94) participants were recruited to generate follow-up questions. All were UG and PG students from a range of science and arts disciplines at the University of Sussex. Each received either £5 or course credits for taking part.

Procedure

Exam paper tasks. The four interviewees arrived at the lab and were provided with paper-based instructions for four tasks, as well as a laptop that they were told belonged to the Head of the School of Psychology. Each of the four tasks concerned a mock exam paper and a real exam paper for the upcoming exam period. In reality, both papers were mock exams from previous years, one of which had been doctored to appear as though it was for the current year. Each task was varied in terms of scope (narrow or broad) and veracity (truth-telling, hiding, or pretending).

Task 1: copy paper. For Task 1, the instructions asked participants to make a copy of a file. In all conditions they were told where they could find the file on the laptop. The truth-telling instructions for this task asked them to make a copy of the file 'C8035 Social Psychology SAMPLE 2017' (hereafter referred to as the 'mock paper'). The hiding instructions were the same but included the further instruction to also make a copy of the file 'C8508 Cognition in Clinical Contexts REAL 2017' (hereafter referred to as the 'real paper'). It was made clear to them that no copies of the real exam paper were supposed to be made. The pretending instructions told the participants not to make a copy of the mock paper. They were provided with the truth-telling instructions and told that their goal would be to convince the interviewer that they had conducted the task. Each task was varied in terms of scope. For Task 1, the broad scope version of the instructions said 'Please make a hard copy of this file. The printer in this room doesn't work, however, you may use any other method you wish to make a hard copy.' For the narrow scope version, specific instructions were given that involved them accessing a printing server on a university website where they could use the experimenter's log-in details to send the file to be printed by a specific printer.

Task 2: send paper electronically. For Task 2, the truth-telling instructions asked the participant to send a copy of the mock paper electronically to the experimenter. The hiding instructions included the additional instruction to also send a copy of the real paper. The pretending instructions told them not to conduct the task but to pretend they had in the subsequent interview. The broad scope version of the instructions read 'Please send this copy electronically using any method or platform you wish'. The four participants all personally knew the experimenter and, therefore, had a number of viable methods for completing this task. The narrow version of the instructions specifically instructed them to send the file via email.

Task 3: deliver paper by hand. For Task 3, the truth-telling instructions required the participant to deliver a hard copy of the mock paper to a specific faculty member at the university, BD. The hiding instructions included the additional instruction to deliver the real paper to a different faculty member at the university, RS. The pretending instructions told them not to conduct the task but to pretend they had in the subsequent interview. The broad scope version of the instructions read 'You will be given a copy of the file. Please place it the envelope and deliver it to BD/RS's office. Use any method you wish to locate the office'. The narrow version had the same instructions except that they were provided with the room numbers for the two faculty members.

Task 4: destroy paper. In Task 4, the truth-telling instructions asked the participants to destroy the hard copy of the mock paper that they had created in Task 1 to 'avoid it falling into the wrong hands'. The hiding instructions asked them to also destroy a copy of the real paper. The pretending instructions told them not to conduct the task but to pretend they had in the subsequent interview. The broad scope version of the instructions said, 'Please destroy the file using any method you wish, as long as the file is fully destroyed (i.e., not just thrown in the bin)'. The narrow scope version said, 'In the psychology school office you will find a shredder (in the kitchen area). Please use this shredder to destroy the file'.

The interviews. The four interviewees were asked to provide a free account of each of the four tasks. On each occasion, the interviewer stated which task they were referring to and then asked the interviewee to describe how they had carried out that task. When they had finished the free recall the interviewer prompted them, asking if there was anything they would like to add. If they indicated that there was nothing to add, the interviewer would, at this stage, introduce his knowledge of the real paper being copied/sent electronically/delivered to RS/destroyed, depending on the task in

question. The interviewer would then ask the interviewee to explain any involvement they had had in that. Each interview was filmed and subsequently edited so that each video contained one interviewee describing their involvement in one task, creating 16 videos in total.

Question generation. The experiment was presented to participants on a computer screen using Qualtrics software in a lab at the University of Sussex. After being presented with the study information, consent form, and instructions, all were asked to provide basic demographic information. The participants were then shown a 10-minute training video that covered the key framework of the CCE interview technique (Ormerod & Dando, 2015). This video proved to be an effective technique for improving the quality of generated questions in Chapter 4. In the main task, participants were shown all 16 of the interview videos in a random order. They were instructed to imagine that they were the interviewer and that their goal was to establish whether the account given by the interviewee in each clip was true or false. To this end they were asked to generate a question after each clip that they felt was creative, would provide some useful information and might challenge the interviewee's account if it was not genuine. After each clip, the software automatically proceeded to the question generation screen, with a box for them to type the question they had generated. They were given 60 seconds to provide their question, at which point the screen automatically moved on to the next video. This procedure was repeated until all 16 videos had been viewed. This process generated a total of 960 questions.

Card sort. The questions generated by the participants were each printed onto card and grouped into individual sets for each of the 16 interview clips. The authors began by taking the 60 questions generated for one video and deciding if each in turn had similar attributes to any previous questions, or whether they were entirely distinct

from previous questions. As such, a number of question categories began to form. At the end of each set of questions, the authors reviewed the categories, making a firm decision as to whether each question had been correctly sorted. Each category was then ascribed a label and a set of characteristics. When this procedure was complete, the same technique was applied to another set of 60 questions, sorting them into the previously described categories or creating new categories if a question did not fit into any of the previously defined groups. At the end of each set of questions, the authors continued to review and refine the categories. This procedure continued until no new categories were identified.

Results

Card Sort 1

For the first set of 60 questions, the card sort identified four distinct categories:

Uncodeable*:* This group constituted questions that were incomprehensible, entirely irrelevant, incomplete or otherwise unable to be meaningfully categorised.

Global veracity checks*:* These were checks of veracity. They essentially asked the interviewee to state whether they were telling the truth or not or whether they might have unwittingly or mistakenly undertaken the wrong task, such as “Do you think someone could have got the instructions wrong?”

Information provision*:* These questions sought further clarification regarding information that the interviewee had provided in their account. They did not present a challenge to the interviewee or progress the investigation beyond the interviewee’s statement. They tended to be verifiable but predictable queries, such as “Where outside the office door did you leave the envelope?”

Information challenges*:* These were questions that challenged the interviewee to reveal necessarily episodic information regarding the task (i.e., information they would possess only if they had actually undertaken the task). They were akin to tests of expected knowledge, potentially verifiable, and less predictable than the other three categories, such as "Describe BD's office door- the colour, any posters or notes on it, etc"

Card Sort 2

The procedure was repeated with another set of 60 questions. This time a distinction was made between two components of the 'information provision' category. The category was subsequently split into the following two distinct categories:

Inconsequential*:* These questions challenged a part of the interviewee's account that was not investigatively relevant and did not progress the investigation in any meaningful way. They were not necessarily predictable but focused on inconsequential motives, such as "did you think twice about leaving the paper outside the office door?"

Information clarification*:* These questions were essentially 'fact-checkers'. Similar to the previously identified 'information provision' category, they tended to be seeking clarification regarding an element of the interviewee's account, but not testing the account in any meaningful way. They were verifiable and relevant but often predictable, such as "what was the email address that you sent the file to?"

Card Sort 3

A third card sort was conducted on the next set of 60 questions. On this occasion the 'information challenges' category was further fragmented into two distinct categories:

Episodic challenges (context): These questions were ones that required the interviewee to reveal episodic knowledge regarding the context or environment associated with the particular task. They tended to be easily verifiable, relevant and less predictable than 'information clarification' questions. They often challenged the interviewee's account, such as "Describe the exact location of the shredder within the school office."

Episodic challenges (action)*:* This category had the same properties as the 'episodic challenge (context)' questions but the focus was instead on the interviewee's actions, such as "Explain why it took three attempts to destroy the paper."

Card Sort 4

During the fourth set of questions, a new category was identified:

Elephant traps*:* These were essentially 'trick questions' designed to catch the interviewee out. They tended to occur when the interviewee had mentioned something that was potentially inconsistent. They were often attempting to test an apparent contradiction, such as "How do you know where RS' (the member of faculty who the real exam paper had been erroneously delivered to) office is?" (given that they had denied delivering the paper to this office).

Card Sort 5 and Onwards

The authors continued to follow the same procedure with the remaining sets of questions. However, the seven previously identified categories were sufficient in capturing each subsequent set and no new categories were identified. Therefore, the final seven categories, established by the card sort, were as follows: Uncodeable, Inconsequential, Global Veracity Check, Information Clarification, Episodic Challenge (context), Episodic Challenge (action), and Elephant Traps.

Distinguishing Dimensions

Having established these seven question types, the authors attempted to discern which properties contributed towards the distinction between high- and low-quality categories. For example, superficially, there is little to distinguish between ‘inconsequential’ or ‘information clarification’ questions and the ‘episodic challenge’ or ‘elephant trap’ questions, yet the latter are instinctively better quality. In order determine potential dimensions on which each category differed, the authors took each set in turn and discussed its properties. Relevance to the investigation was the initial dimension to be determined. It was clear, for example, that the key difference between Inconsequential questions and Global Veracity Checks was that the Global Veracity Checks were relevant to the enquiry and Inconsequential questions were not. Likewise, investigate relevance was key in distinguishing between the two Episodic Challenge types, in that the actions involved in an incident are more investigatively relevant than the context. In turn, context related Episodic Challenges are still more investigatively relevant than Inconsequential questions.

The second dimension identified was predictability. Asking questions that a deceptive interviewee has not anticipated removes the option for them to present a pre-prepared lie, raising their cognitive load in turn (Vrij et al., 2009). This dimension was also applicable in distinguishing between the question types identified in the card sort. For example, Information Clarification questions, Episodic Challenge (action) questions, and Elephant Traps all tend to be highly relevant but Episodic Challenge (action) questions tend to be more unpredictable than Information Clarification questions and, in turn, Elephant Traps tend to be more unpredictable than Episodic Challenge (action) questions.

The final dimension identified was the type of knowledge required to answer the question. Sporer's (2016) work on detecting deception highlights the important distinction between episodic and semantic knowledge in relation to positive interview outcomes. Here we identified that a distinction could be made between questions that could be answered from a general, semantic knowledge of the situation or event, and questions that would require the interviewee to have really experienced that event in order to provide an answer. For example, Global Veracity Checks and Information Clarification questions are both high in Relevance and low in Unpredictability but the Information Clarification questions tend to require more first-hand experience of the event in order to provide an answer. In turn, The Episodic Challenge questions require a higher level of episodic knowledge than the Information Clarification questions.

As such, a system was developed whereby each of the identified question types were assigned as being low, medium or high in terms of these three dimensions. The Uncodeable category was not included in this analysis as these questions were of no value. As can be seen in Table 1, the three dimensions were capable of separating the six question types.

Table 1

Difference between six identified question types in terms of the three proposed dimensions of question quality.

	Relevance	Unpredictability	Episodic Knowledge Required
Inconsequential	Low	Low	Low
Global Veracity Checks	High	Low	Low
Information Clarification	High	Low	Medium
Episodic Challenge (context)	Medium	Medium	High
Episodic Challenge (action)	High	Medium	High
Elephant Traps	High	High	High

Discussion

The pilot study used a qualitative, bottom-up approach to explore an objective measure of question quality. In Chapter 4, questions were rated according to the standard definition of creativity. However, this proved to be an unreliable method. The pilot study presented here gathered a large sample of test questions, generated by novices, and employed a card sort technique to discern the factors that distinguish good quality questions from poor quality questions. The card sort indicated seven distinct categories of question type. Subsequently, three over-arching dimensions, encapsulating distinctions between the seven categories, were identified that theoretically should be capable of distinguishing between high-quality and low-quality questions. The three dimensions were investigative relevance, unpredictability, and the type of knowledge probed. According to this theory, a question that is relevant to the incident under investigation, unpredictable, and requiring a response that focuses on episodic knowledge, should be a good quality question.

There are empirical findings that support this theory. Investigative relevance is arguably the most intuitive of the three components. Essentially, it pertains to the utility of the question, in so much that a question that focuses on investigatively irrelevant details is very unlikely to result in any useful information being obtained. The PEACE model of interviewing, which all police officers in the UK are trained in, encourages interviewers to plan and prepare for interviews they are due to carry out, establishing the specific, relevant information necessary to elicit a reliable account of the incident in question. The unpredictability component is supported by Vrij and colleagues' (2009) work on the UQ approach. The experiment presented in Chapter 3 brought this technique into question as an isolated component. However, there is evidence to suggest that unpredictability is a useful tool when incorporated within a wider framework, such as the CCE technique (Ormerod & Dando, 2015). Finally, the advantages of questioning on episodic information over semantic information has been outlined previously in this thesis. Recalling episodic information is more cognitively demanding than semantic information (Taylor & Dando, 2018). Lying is already a more cognitively demanding task than telling the truth (Debey et al., 2015; Vrij et al., 2006). Therefore, increasing the mental effort required by a deceptive interviewee, by asking for episodic information, should result in more pronounced differences between the verbal behaviour of truth-tellers and liars (Dando et al., 2015). As such, a question which investigates an interviewee's episodic experience of an event should have more value than one which simply seeks semantic information.

Whilst there is empirical support for the three identified components detailed here, at present their value is theoretical. Furthermore, the card sort methodology used to develop the three dimensions is arguably subjective in nature and has the potential to be constrained by the researchers' academic backgrounds. For example, they may lean

towards the fundamentals of the CCE framework (Ormerod & Dando, 2015). Therefore, it is necessary to empirically test the 3-dimensional rating scale in order to determine its objectivity, its reliability as a scale, and its value in determining the quality of test questions.

Main Experiment

Introduction

The interview videos created in the pilot study were shown to a new group of both novice and expert participants, who were each asked to generate a set of follow-up test questions. Four expert judges rated the questions in terms of the three factors identified by the card sort – relevance, unpredictability and type of knowledge required. Given the findings of Chapter 4, the expert judges were also asked to rate each question for general quality. Additionally, a group of novice judges rated the questions for general quality. Therefore, the new 3-dimensional rating model was tested in two ways: firstly, looking at the inter-rater reliability of the four experts' ratings and, secondly, by measuring the extent to which the ratings on this scale map on to the more subjective quality ratings made by the same experts. If the 3-dimensional model is effective in capturing the differences between good- and poor-quality questions, one would expect to see a positive correlation between ratings made by individual experts on the two rating methods.

In terms of the relationship between expert and novice ratings, previous research would suggest that there is potential for there to be disagreement between the two groups. There is research, across a wide range of disciplines, to suggest there may be some overlap between expert and novice ratings. For example, Plucker, Kaufman, Temple, and Qian (2009) showed there was a moderate correlation between experts'

and novices' ratings of films. However, Hekkert and Van Wieringen (1996) showed that, whilst experts and novices agreed on the originality of a series of art works, they disagreed in terms of craftsmanship and quality. Moreover, other studies have found there to be no overlap at all, for example, Dorfman (1996) found no agreement between experts and novices when rating the quality of fictional writing, whilst Runco and colleagues (1994) also showed disagreement between experts' and novices' ratings of art work, arguing that experts tend to be overly critical and less sensitive to differences in ability.

The tasks carried out by the four interviewees in the pilot studied were varied in terms of scope and veracity. The reason for varying scope was based on the findings of Chapter 4. The results of that experiment revealed that the quality of follow-up questions generated by participants was affected by the temporal perspective of interview topic. We theorised that this was due to differences in the scope of episodic information inherently available to the interviewer, with a broad scope widening the episodic information available to draw on and in turn resulting in higher quality questions. This idea is supported by Sporer's (2016) invocation of a schema theory of working memory which states that recollection of genuine experiences will have more unique details available than imagined experiences, that is, topics that grant an interviewer greater opportunity to enquire about events unique to the interviewee's experience should be more useful to the investigation.

Veracity was manipulated so that the interviewees carried out two tasks truthfully and two tasks deceptively. Furthermore, a distinction was made between two forms of deception: hiding and pretending. There is a wealth of research to show the difference between truth-tellers' and liars' verbal behaviour (e.g., Bogaard, Colwell, & Crans, 2019; Hartwig et al., 2007; Vrij, Leal, & Fisher, 2018). However, there is an

important, but often overlooked, distinction to be made between different forms of lying. Lying to hide an event requires the construction and maintenance of two sets of world models: the one that you wish to portray and the real world you wish to hide. This duality is known to require a greater level of cognitive effort (Ormerod & Richardson, 2003). Whereas, lying in order to pretend that a series of events have taken place when they have not, requires only a single portrayed world and can often be embedded within genuine personal experiences. Memon and colleagues' (2013) unpublished literature review showed the effect that this difference in lying can have on interview outcomes. Reality monitoring analyses revealed that liars and truth-tellers exhibit greater differences in verbal behaviour when lying involves hiding an act rather than pretending to have conducted an act. As such, in relation to the present study, differences in the verbal content of interviewees' responses to hide or pretend events may elicit differences in quality between the follow-up question generated by participants.

The experience of question generators was also investigated in the present experiment. In Chapter 4, the results indicated that a short training video improved the performance of novice participants. However, the mean quality rating for their questions was still below the mid-point of the scale, suggesting that novices, in general, are not able to generate good quality test questions. This presents a problem for the current research, as investigation into the creative process of question generation is somewhat limited if the sample does not contain a reasonable quantity of good quality questions. There is no research looking into expert/novice differences in question generation. However, there is research from other areas that shows that experts tend to demonstrate wider domain-specific knowledge, possess greater perceptual skills and are capable of applying more complex thought processes to problem-solving than novices (Klein & Hoffman, 1992; Mosier, Fischer, Hoffman, & Klein, 2018). In a more investigatively-

relevant field, Fahsing and Ask (2016) showed trained police detectives and novice police officers fictional accounts of two missing persons cases and asked them to generate as many investigatively relevant hypotheses and actions as they could. The results showed that, for UK-based participants, the expert detectives generated more alternative hypotheses and actions than the novice participants. Given these findings, it is reasonable to assume that trained investigators in the present study would generate better quality questions than novices.

Investigating the effects of expertise raises an additional point to consider: what constitutes expertise? Of course, there are numerous ways to define expertise, making the term somewhat vague. In the case of investigative interviewing, one might consider a person who conducts such interviews as part of their job (e.g., police officers) to be an expert. However, most police officers have only received the week-long Level 1 PEACE training, and evidence suggests that some of the skills learnt in that week are soon forgotten for some, with the training having little effect on their subsequent interviewing practices (Clarke et al., 2011). Some officers, who are responsible for conducting interviews concerning more serious crimes may have received the more substantial 3-week-long Advanced Interview Training (Level 3 of the PEACE model). Given the thoroughness of this training and the real-world experience held by these individuals, the term ‘expert’ may seem fitting, but it is still not possible to effectively determine their expertise from their level of training and job title alone. This presents a problem for the present study, with regards to looking at differences between novice- and expert-generated questions. Therefore, to allow for efficient data collection, the term ‘expert’ was defined here as any individual with five or more years’ experience in a job role that requires investigative interviewing. Whilst the participants’ expertise may be brought into question, their experience in the task at hand should still afford them an

advantage over the ‘novice’ participants, who have never performed investigative interviewing.

The present experiment was designed, firstly, to examine the extent to which the three dimensions of question quality, identified in the pilot study, are able to predict global judgements of question quality and, therefore, provide a potential model for the design of good investigative questions. Secondly, the experiment explored the effects of scope of available episodic information, veracity of interviewee, and experience of question generator on the quality of question generation. Based on the card sort conducted in the pilot study, we expected to find that experts’ scores on the 3-dimension rating scheme would show acceptable levels of inter-rater reliability (Hypothesis 1) and correlate with their own ratings of general quality (Hypothesis 2). Based on findings across a broad spectrum of domains (e.g., Dorfman, 1996; Hekkert & Van Wieringen, 1996; Plucker et al., 2009) we expected to find little correlation between experts’ ratings and novice ratings of quality (Hypothesis 3). Runco and colleagues (1994) argue that experts tend to be overly critical and less sensitive to ability than novices when judging an item’s value. Therefore, we expected to find that the novice judges rate the questions higher than the experts, as well as making a clearer distinction between the novice and expert generated questions (Hypothesis 4). However, based on Fahsing and Ask's (2016) work, expert question generators should produce higher quality questions than novices in general (Hypothesis 5). The theory put forward to explain the findings of Chapter 4 was that the quality of generated questions will be affected by the scope of episodic information available within the context of the interview. As such it was predicted that tasks with a broad scope would result in higher quality questions than tasks with a narrow scope (Hypothesis 6). Finally, given the potential difference in cognitive load imposed by lying to hide an act and lying to pretend an act has taken

place (Memon et al., 2013; Ormerod & Richardson, 2003), we expected to find that there is a difference in the quality of questions generated when the interviewee is hiding and when they are pretending (Hypothesis 7).

Method

Participants.

Question generators.

Novices. Thirty-six females (M_{age} = 19.31, SD = 1.06) and four males (M_{age} = 20.00, SD = 1.83) were asked to generate follow-up questions for the study. Participants were UG and PG students from a range of science and arts disciplines at the University of Sussex. Each received either £5 or course credits for taking part in the study.

Experts. Three female (M_{age} = 53.33, SD = 14.57) and nine male (M_{age} = 46.56, SD = 10.33) experts were also asked to generate follow-up questions for the study. To be included, they were required to have a minimum of five years' experience in a role that involved investigative interviewing. Experience ranged from 8 to 34 years (M = 17.92, SD = 7.86). They were asked to state their current job role (or last investigative role held if retired). The following roles were self-identified by participants: police officer (2), detective constable, information assurance manager, detective sergeant (2), police inspector, principal environmental health officer, detective chief inspector and probation officer (3).

Question raters.

Novices. Seventy-two females (M_{age} = 24.68, SD = 4.50) and eight males (M_{age} = 28.38, SD = 6.32) provided general quality ratings for the study. Participants were all students or faculty members at the University of Sussex. None had any previous experience of interviewing. Each were paid £5 for taking part in the study.

Experts. Two female and two male interviewing experts were recruited to rate the creativity of the questions generated. Two of these were from an academic background: R1 was aged 58 and had 10 years of interviewing experience at Professionalising Investigations Programme Level 3; R2 was aged 33 and had 9 years of interviewing experience having completed advanced Achieving Best Evidence training. The remaining two were from a law enforcement background: R3 was aged 54 and had over 25 years of experience as a Tier 5 home office interview advisor; R4 was aged 56 and had over 35 years of experience as a criminal justice and investigative interview consultant.

Design

A mixed design was employed in the interview videos. Each interviewee completed each of the four tasks: copy paper (n = 4), send paper electronically (n = 4), deliver paper by hand (n = 4), and destroy paper (n = 4). The tasks were varied in terms of scope, with a narrow scope version (n = 8) and a broad scope version (n = 8) version of each task. Finally, the interviewees were randomly assigned to carry out each task in one of three ways: truthfully (n = 8), hiding (n = 4) or pretending (n = 4).

A mixed design was also employed for the question generating. There were two independent groups that took part: experts (n = 12) and novices (n = 40). All participants were randomly assigned two of the four interviewees' videos. In total, each participant viewed eight videos and provided a test question after each.

Procedure

Question generation. The interview videos created in the pilot study were employed in the main study. Both groups of participants (experts and novices) were randomly allocated two of the four interviewees' videos and were shown all four of the

clips for each of those interviewees. They were instructed to imagine that they were the interviewer and that their goal was to establish whether the account given by the interviewee in each clip was true or false. To this end they were asked to generate a question after each clip that they felt was creative, would provide some useful information and might challenge the interviewee's account if it was not genuine.

The novice participants were shown the 10-minute training video, based on the CCE interview technique (Ormerod & Dando, 2015), developed in Chapter 4. The expert participants were not provided with this training video as it was assumed that their experience in investigative interviewing would supersede the advice it conveyed.

The experiment was presented to participants on a computer screen using Qualtrics software. Novices completed the study in a lab at the University of Sussex, whereas the experts completed the study remotely. After being presented with the study information, consent form, instructions and the training video (in the case of the novice participants), all were asked to provide basic demographic information and the experts were asked to provide further details of their investigative experience. In the main task, the videos were shown in task order so that the narrative was logical. After each clip, the software automatically proceeded to the question generation screen, with a box for them to type the question they had generated. They were given 60 seconds to do this, at which point the screen automatically moved on to the next video. This procedure was repeated until all eight videos had been viewed.

This generated a total of 416 questions. However, there were a number of duplicate questions. Questions were considered to be duplicates when they were attempting to gather the same information, even if distinct language was used. For example, in response to one of the 'Destroy Paper' tasks, the questions "can you

remember who it was that showed you how to shred it?" and "Who showed you how to shred it in the office?" were considered to be duplicates as they were both asking for the name of a specific person mentioned in the interviewee's account. However, the question "can you describe the person who you asked to help you with the shredder?" was not considered to be a duplicate in this set as it was asking for a description of the person, as opposed to simply their name.

The lead author categorised the duplicate questions for each of the 16 videos and then randomly chose one of each set of duplicates to be included in the rating section of the study. In total, 282 of the 416 questions were included in the rating section of the experiment. Before the 282 questions were rated, they were processed so that any leading or closed questions were rephrased as either focussed or open questions. The information sought by the question always remained the same. However, the wording was changed in order to avoid the raters harshly judging the questions based on their form rather than the information they conveyed (Dando, Geiselman, MacLeod, & Griffiths, 2016; Milne & Bull, 2016; Milne, Griffiths, Clarke, & Dando, 2019). For example, in relation to one of the 'Destroy Paper' tasks, the question "Can you remember any details about the mock exam paper?" was reworded to "Describe any details you remember about the mock exam paper." Inter-rater reliability analyses of the rating scales were conducted on the 282 non-duplicate questions that had been rated. However, for subsequent analyses the remaining duplicates in each set were allocated the same rating as the one chosen to be rated in order to maintain equal group numbers.

Question rating.

General quality. Both novice and expert participants provided general quality ratings for the study, using Qualtrics software. Novices completed the ratings in a lab at

the University of Sussex, whilst the experts completed the task remotely. All were asked to provide basic demographic information and experts were further asked to detail their interviewing experience. They were then provided with the instructions for the ratings, which included information about the four tasks completed by the interviewees. The experts were shown the four videos associated with each interviewee (the order of interviewee was randomly varied), followed by all of the questions generated for each clip, presented one at a time. They were asked to rate each question on a 1-7 scale of general quality. This was repeated until all 282 questions had been rated. For the novices, the procedure was the same, except they only rated the questions generated for one of the interviewee's four videos.

Three-dimensional model. In order to explore the three-dimensional model developed in the pilot study, the four expert raters completed the same procedure as outlined for the general quality ratings, though this time rating each question for 'Relevance', 'Unpredictability' and 'Knowledge'. Definitions were provided for each. Relevance was explained as follows: 'is the question focused on details that are **relevant** to those contained in the interviewee's statement, or is it more concerned with **irrelevant** details?' Unpredictability was explained as follows: 'is the question asking for obvious details that the interviewee may find **predictable** or is it asking for less obvious details that may be considered **unpredictable**?' Finally, Knowledge was explained as such: 'is the type of **knowledge** required specific to the interviewee's experience (i.e., **episodic**) or is it something that does not require specific experience of the event (i.e., **general**)?' Each dimension was rated on a 1-10 visual analogue scale. Labels were provided at each end of the three scales. For Relevance the scale was Irrelevant (0) to Relevant (10); for Unpredictability the scale was Predictable (0) to Unpredictable (10); and for Knowledge the scale was General (0) to Episodic (10).

Intraclass Correlation Coefficient Analysis

Inter-rater reliability for each component of the 3-dimensional model and general quality ratings was calculated using Intraclass Correlation Coefficients (ICC). In each instance a two-way random-effect model based on average ratings and consistency (ICC(C,k)) was applied, as the intention was to calculate the mean of the judges' ratings for each scale. As outline in Chapter 4, we applied Cicchetti and Sparrow (1990) guidelines which suggest that scores above .70 represent an acceptable level of agreement and employed this as a threshold for inter-rater reliability in the present study.

Results

Inter-Rater Reliability

3-dimensional scale. Three initial ICC(C,k) analyses were conducted on the three components of the scale. For the Relevance component, the inter-rater reliability was .61, 95% CI [.57, .71]. For the Unpredictability component, the inter-rater reliability was .65, 95% CI [.59, .72]. Finally, for the Knowledge component, the ICC was .80, 95% CI [.76, .84]. Whilst Relevance and Unpredictability fell below the threshold, both represented moderate-to-good reliability depending on chosen guidelines (Cicchetti, 1994; Koo & Li, 2016). Ratings of the Knowledge component did meet the threshold. In order to examine the reliability of the scale as a whole, the sum of each expert's ratings across the three components was calculated for each question and a fourth ICC(C,k) analysis conducted. This showed the reliability of the scale overall to be .76, 95% CI [.71, .80]. This exceeds the threshold and can be interpreted as good-to-excellent. Given the novelty of the scale, it was decided to continue with further exploratory analysis using the summed scores of the three components.

General quality ratings.

Experts. The four experts rated each question for quality on a 1-7 Likert scale. An ICC(C,k) analysis was conducted and revealed that reliability was .51, 95% CI [.41, .60]. This is at the low end of moderate-to-good and does not meet our threshold for acceptable reliability. In order to test certain hypotheses, further analysis was conducted using an aggregated average of these scores. However, such analyses must be treated with caution given the poor inter-rater reliability. As the same four experts were employed to conduct both sets of ratings, it was possible to look at correlations between the scales for each of their individual ratings.

Novices. General quality ratings were also gathered from 80 novice raters, each judging one of the four interviewee's set of clips (i.e., there were 20 ratings for each video clip in total). A good-to-excellent inter-rater reliability was found between the participants. The ICC(C,k) was .83, 95% CI [.79, .85]. Therefore, the mean of the 20 ratings was taken for each question and used in subsequent analysis.

Correlations Between Scales

Individual experts. There was a strong correlation between the mean of the four experts' ratings on the 3-dimensional scale and the mean of their general quality ratings, ($r = .61$, $n = 410$, $p < .001$). However, the expert ratings of general quality did not meet the threshold set for reliability, and therefore, this correlation should be interpreted with caution. However, it is possible to interpret the correlations between their individual ratings.

Table 2 shows that for two of the experts, Knowledge was most highly correlated with their quality scores, whilst Relevance was most highly correlated for the other two. Unpredictability showed a weak correlation with general quality for all four

of the raters, suggesting that none of them considered the unpredictability of the questions to be indicative of quality. Overall, there was a medium or large positive correlation between the overall summed scores on the 3-dimensional model and the general quality scores for three of the judges. However, there was no such correlation found for Rater 1's scores.

Table 2

Correlation between each raters' general quality rating and their ratings on the 3-dimensional model.

	Relevance	Unpredictability	Knowledge	Total[1]
Rater 1	.17	-.03	.04	.09
Rater 2	.27	.17	.56	.55
Rater 3	-.12	.22	.34	.36
Rater 4	.50	-.16	.31	.40

Note: [1] Total refers to the sum of the three components. n = 410 for all correlations.

Experts and novices. In order to determine whether the 3-dimensional ratings made by the experts mapped on to the novices' ratings of general quality, a Pearson's correlation was conducted. There was no relationship between the mean of the novices' general quality ratings and the mean of experts' 3-dimensional ratings, (r = .08, n = 410, p = .109). This suggests that the ratings made by experts using the 3-dimensional model did not map onto ratings of general quality made by novices. Moreover, there was only a weak correlation between the experts' and novices' general quality ratings, (r = .26, n =410, p < .001). This finding should be treated with caution due to the poor inter-rater reliability within the experts' ratings. However, it further suggests that experts' and novices' opinions differed on the quality of test questions.

Difference in Experts' and Novices' Ratings

In order to investigate the difference between experts' and novices' average general quality ratings a repeated measures t-test was conducted. Novices, in general, rated the quality of the questions significantly higher (M = 4.71, SD = .82) than the experts (M = 3.58, SD = 1.02), $t(409) = 20.33$, $p < .001$, $d = 1.22$, 95% CI [1.01, 1.43]. However, this analysis must be treated with caution given the low inter-rater reliability exhibited within the experts' quality ratings.

Effect of Expertise, Scope and Veracity

Experts raters. In order to determine the effect of experience of the question generators, scope of the interview topic and veracity of the interviewee, a 2 (Experience: expert vs. novice) × 2 (Scope: narrow vs. broad) × 3 (Veracity: truth-telling, hiding, pretending) ANOVA was conducted on the mean 3-dimensional ratings made by experts. The results showed that there was not a significant main effect of Experience, $F(1, 270) = 0.54$, $p = .46$, or Veracity, $F(2, 270) = 0.71$, $p = .49$. Nor were there any significant interactions. However, there was a significant main effect of Scope, $F(1, 270) = 5.22$, $p = .02$, $\eta_p^2 = .02$. A follow-up t-test revealed that interviews with a narrow scope (M = 17.04, SD = 3.52) resulted in significantly higher average ratings than those with broad scope (M = 15.70, SD = 4.15), $t(280) = 2.92$, $p = .004$.

This finding went against prediction. In order to seek an explanation for this finding, Scope was further analysed within the context of the three components that make up the model. Figure 1 shows the difference between scores assigned to questions generated for narrow scope interviews and broad scope interviews, on each of the three components.

A series of independent t-tests were carried out to examine the effect of scope. For relevance ratings, there was no significant difference between narrow scope interviews (M = 6.53, SD = 1.58) and broad scope interviews (M = 6.30, SD = 1.79), $t(408) = 1.40$, $p = .16$. For the unpredictability ratings, there was also no significant difference between narrow scope interviews (M = 4.69, SD = 1.66) and broad scope interviews (M = 4.51, SD = 1.71), $t(408) = 1.12$, $p = .26$. However, there was a significant difference between narrow scope interviews (M = 6.18, SD = 2.50) and broad scope interviews (M = 5.37, SD = 2.88) on the knowledge ratings, $t(408) = 3.07$, $p = .002$, $d = 0.30$, 95% CI [0.03, 0.58]. This suggests that when the scope of the interview was narrow, participants generated questions that required more episodic knowledge than when the scope of the interview was broad.

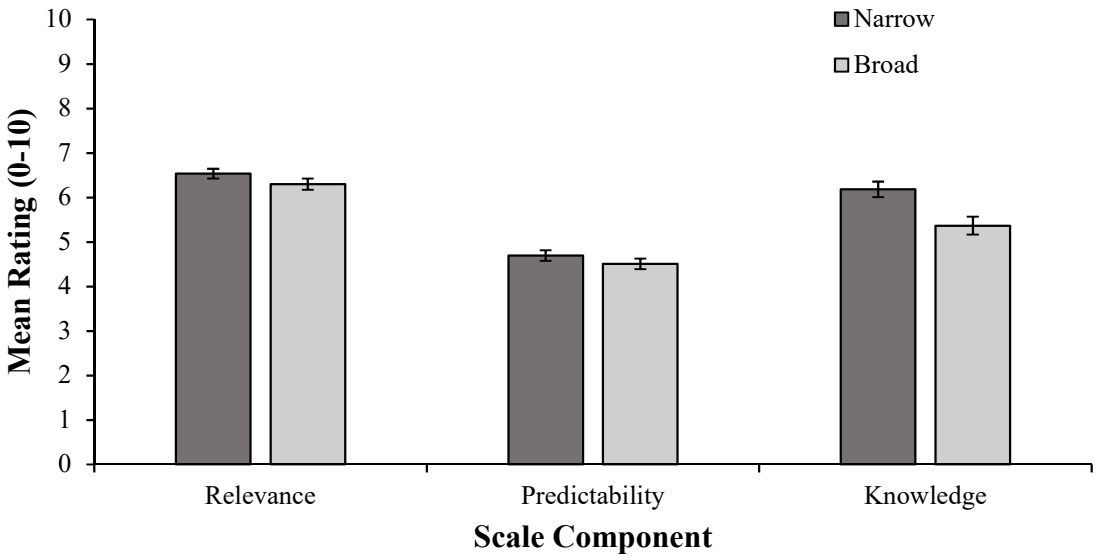

Figure 1. The effect of scope of interview on the mean ratings of relevance, predictability and knowledge. Error bars represent +/- 1 SEM.

Novice raters. In order to determine the effect of experience of the question generators, scope of the interview topic and veracity of the interviewee, a 2

(Experience: expert vs. novice) × 2 (Scope: narrow vs. broad) × 3 (Veracity: truth-telling, hiding, pretending) ANOVA was conducted on the mean general quality ratings made by the novices. In this instance, the results showed the converse pattern of the experts' ratings. There was no significant main effect of Scope, $F(1, 410) = 1.10$, $p = .30$. However, there was a significant main effect of both Experience, $F(1, 410) = 4.52$, $p = .03$, $\eta_p^2 = .01$, and Veracity, $F(2, 410) = 6.42$, $p = .002$, $\eta_p^2 = .03$. There were no significant interactions.

To further examine the effect of experience, a follow-up t-test revealed that novices rated the quality of the question generated by experts (M = 4.89, SD = 0.77) significantly higher than the questions generated by the novices (M = 4.66, SD = 0.83), $t(408) = 2.40$, $p = .02$ $d = 0.28$, 95% CI [0.05, 0.51]. To examine the effect of veracity, follow-up planned contrasts revealed that quality ratings were not significantly different for the questions generated when interviewees were lying (i.e. hiding and pretending combined; M = 4.70, SD = 0.80) and when interviewees were telling the truth (M = 4.73, SD = 0.85), $t(408) = 0.43$, $p = .67$. However, quality ratings were significantly higher when the interviewee was hiding (M = 4.93, SD = 0.68) than when they were pretending (M = 4.46, SD = 0.85), $t(203) = 4.42$, $p < .001$, $d = 0.61$, 95% CI [0.33, 0.89].

Discussion

In the main study, an experiment was designed with which to test the 3-dimensional rating model developed in the pilot study. Additionally, the effects of scope, veracity and experience were investigated. In general, the findings were mixed. The 3-dimensional rating model proved to be a reliable measure when the three components were summed together. Additionally, for three of the raters there was a

positive correlation found between their ratings using the 3-dimensional model and the more subjective general quality scale. This provides tentative support for the efficacy of the new rating scheme. The results suggested that experts and novices have differing opinions on the value of investigative questions, with novices rating the questions higher than the experts. Moreover, the novice raters were more sensitive to the experience of the question generators than the expert raters. The theory developed in Chapter 4 that the scope of available episodic information would affect the quality of subsequently generated questions was confirmed, however, in the opposite direction than was predicted. There was an effect of scope on the experts' ratings, however, follow-up tests revealed that better quality questions were generated when the tasks had a narrow scope, contrary to prediction. Finally, the findings of the novice ratings revealed that when the interviewee was lying to hide the truth, better quality questions were generated than when the interviewee was lying to pretend an incident had taken place. Overall, the findings shed some light on the nature of question generation, though much investigation is still required to explore the process further.

Given the potentially subjective nature of the card sort that led to the development of the scale, it was of utmost importance to establish, empirically, whether the 3-dimensional model was objective and reliable. The findings suggested that there was mixed support for this. For the individual dimensions, the results indicated that the inter-rater reliability of the Relevance and Predictability components was moderate-to-good (Cicchetti, 1994; Koo & Li, 2016). However, both fell below the threshold set in advance. Despite this, the Knowledge dimension did exhibit a high level of agreement, meeting the acceptable threshold. Moreover, when the three components were summed to an overall measure, the inter-rater reliability was good-to-excellent and exceeded the

threshold. This provided tentative support for Hypothesis 1 and gave us some confidence in the efficacy of the model.

Looking at the correlations between overall ratings on the 3-dimensional model and the general quality scale for each individual rater, also provided mixed support for Hypothesis 2. Two of the raters' scores showed a medium positive correlation, whilst another had a strong positive correlation. However, one rater's scores were not correlated at all. In terms of the individual dimensions of the scale, the results suggested that Relevance and Knowledge mapped onto quality ratings more so than Unpredictability. In Chapter 3, the use of unanticipated questions was brought into question and in Chapter 4, very weak correlations were found between judges' ratings of novelty and quality. Taken together, these findings bring the value of unpredictability into doubt. However, the findings of Vrij and colleagues' (2009) and the fact that unpredictability is one of the key components of the CCE technique (Ormerod & Dando, 2015) arguably suggest that this dimension should continue to be included until the model has been more thoroughly tested.

There is much research across a broad spectrum of domains which suggests that there is little to no overlap between quality judgements made by experts and novices (e.g., Dorfman, 1996; Hekkert & Van Wieringen, 1996; Plucker et al., 2009). The findings of the present study support this. There was no correlation between the experts' overall ratings on the 3-dimensional model and the novices' quality ratings, whilst there was only a weak correlation between the experts' and novices' quality ratings, supporting Hypothesis 3. Furthermore, novice ratings were higher overall than the expert ratings, and also distinguished between expert- and novice-generated questions, which expert ratings did not. The supports Hypothesis 4, as well as the suggestion made

by Runco and colleagues (1994) that experts tend to be more critical and less sensitive to ability than novices.

Despite these findings, some research suggests that the difference between novice and expert ratings may be due to the way that the different groups categorise the objects that are to be rated. Novices tend to categorise superficially, based on commonalities in object appearance or structure; whereas experts tend to categorise conceptually, based on their long-term experience of that domain (McKeithen, Reitman, Rueter, & Hirtle, 1981). So, whilst the novice ratings in the present study discriminated between novice and expert question generators, that does not necessarily imply that the experts' questions were superior in a way that was apparently unclear to the expert raters. It may mean that the experts' questions were simply similar in a way that appeared superficially superior to the novice raters.

The results did support the hypothesis that novice raters would be more sensitive to differences between the experience of question generators than experts. However, it was still expected that experience would influence both sets of ratings. The findings indicated that this was only the case for novice ratings, with no significant main effect of expertise found for the experts' ratings, providing only mixed support for Hypothesis 5. Fahsing and Ask's (2016) study provided support for the intuitive notion that experts in a particular domain will generate higher quality products relevant to that domain. Even within the novice ratings, which demonstrated a difference in quality between novice- and expert-generated questions, the expert-generated questions' average rating was still around the mid-point of scale.

Experts were included in the present study in order to increase the quality of the generated questions, allowing for a more in-depth assessment of the factors that

contribute towards quality. The findings indicate that question generation, even among trained investigative professionals, is an inherently difficult task. However, it should be noted that the term expert was defined as any individual with five or more years' experience in an investigative role. This guarantees experience in the required task, but not necessarily expertise. This is a limitation of the present study and, as such, the findings must be treated with caution as the lack of quality observed in the experts' generated questions may in fact be as a result of their lack of expertise. This indicates the need for continued exploration of expertise, in relation to question generation. Future studies may wish to be more prescriptive with regards to the definition of the term 'expert', perhaps by conducting a priori competence tests, or by examining genuine interviews conducted by Tier 3-trained investigators.

In Chapter 4, the findings revealed an interaction between the topic of interview and the temporal context of the topic on subsequent question generation ability. The theory provided to explain this finding was that the scope of available episodic information, inherent to the topic in hand, will affect the ability to generate good quality test questions. We argued that the more broad the scope, the more opportunity there is to generate good quality questions. The findings of the present study refute this theory. There was no effect of scope on the novice ratings. There was an effect of scope on the experts' ratings, however, simple follow-up tests revealed this effect to be in the opposite direct than predicted. Questions that were generated from narrow scope tasks were rated higher than those generated from the broad scope tasks. Moreover, further analysis revealed that the effect appeared to be specific to the Knowledge dimension of the 3-dimensional model. There was no difference in rating between narrow and broad scope tasks for Relevance or Unpredictability, however, there was a difference for

Knowledge probed, with questions being rated as probing more episodic knowledge when the task had a narrow scope than when the scope was broad.

This finding contradicts the findings of Chapter 4. However, the explanation might be found in the difference between the types of interviews used in each study. In Chapter 4, the interviews were not forensically motivated. They consisted of relaxed conversations concerning general topics, such as an individual's hobbies or their employment. In the present study, the interviews were designed to be more forensic in nature. They concerned a series of specific incidents, during which an illicit activity had taken place. Perhaps the difference in tone between these two forms of interview can explain the findings regarding scope. It may be the case that having a broad scope is beneficial when an interviewer is discussing general topics, not specific to a particular incident. However, when a specific incident is under investigation, the results here suggest that a narrow scope is beneficial, perhaps by providing a more finely-tuned framework with which to probe specific episodic details.

The experiment also manipulated the veracity of the interviewees. The findings suggest that expert raters were insensitive to differences in veracity. However, there was an effect on the novices' ratings. Planned contrasts revealed that there was no difference in quality between questions generated for interviews where the interviewee was telling the truth and when they were lying. However, there was a difference in quality between the two forms of lying. Higher quality questions were generated when the interviewee was lying to hide the truth than when they were lying to pretend an event had taken place. This provides support for Hypothesis 7, as well as the cognitive load theory put forward by Memon and colleagues (2013). They argue that hiding is a more cognitive demanding task than pretending, given that the hider must hold two mental representations simultaneously, whilst pretenders do not. In turn, this increase in

cognitive load should be evident in their verbal behaviour (Bogaard et al., 2019). This could account for the difference in quality observed in the present study. The increase in cognitive load, and resulting change in verbal behaviour, associated with hiding, may have provided the question generators with greater opportunity to generate good quality questions.

The tasks devised for the present study were designed in order to have control over scope and veracity within a forensic context. However, there were limitations with the design. Firstly, the tasks were quite simple and did not provide the question generators with a great deal of variety in terms of lines of enquiry. This may have contributed towards the low overall quality ratings, and the findings regarding episodic scope. Secondly, whilst the interviews were more forensic in nature than those conducted in Chapter 4, they still did not reflect a real-world investigation. The illicit element of the task (using the 'real' exam paper) was relatively arbitrary and did not result in a high-stakes situation. This was necessary in order to manipulate the variables under investigation and to avoid the effects of confounding variables. However, in order to further investigate the efficacy of the 3-dimension model, as well as continuing to explore the factors that contribute towards good quality investigative questioning, it will be necessary to apply the model to a set of interview questions taken from a more professional, forensic context, ideally with genuine outcome measures. If questions which rate high on the 3-dimensional model are shown to result in positive interview outcomes, in genuine forensic interviews, this would increase confidence in the efficacy of the scale and be a step towards providing the investigative community with a useful measurement tool.

General Discussion

The two-part experiment described in this paper was designed in order to identify a novel method for rating the quality of investigative interview questions and to explore potential factors and conditions associated with generating such questions. The pilot study used a card sort to identify seven distinct question types, which were further refined down to three dimensions of question quality: Relevance, Unpredictability and Knowledge probed. This novel rating model was tested in the main experiment. The results provide tentative support for the reliability of the 3-dimensional model. However, further empirical investigation is required in order to determine its efficacy in a more forensic setting.

The findings supported the suggestion that novice and expert raters tend to disagree in their quality ratings (Dorfman, 1996; Runco et al., 1994). There was no correlation between the experts' ratings on the 3-dimensional model and the novices' general quality ratings. Furthermore, the novice's rated the questions higher overall than the experts' and distinguished between expert and novice question generators, suggesting they are more sensitive to expertise than expert raters.

The expert judges' ratings were affected by the scope of the task. Against prediction, it was found that narrow scope tasks resulted in higher rated questions than the broad scope tasks. This is contrary to the findings of Chapter 4. We offer the explanation that scope has differential effects depending on the context of the interview. When the interview is focused on general, non-specific topics a broad scope might be beneficial. However, when the interview is focused on a specific incident, a narrow scope may provide an interviewer with a clear framework with which to address specific episodic details. Additionally, the novice judges' ratings revealed a difference

in quality between questions that followed a task where the interviewee was lying to hide the truth and where the interviewee was lying to pretend an event had taken place. This provides presents further insight into the factors that contribute towards good quality interview questioning. However, a great deal more investigation is required to explore these factors further.

Overall, the study provides some intriguing clues regarding the creative process of question generation. Scope appears to be context dependent; veracity and, more specifically, the type deception, appears to influence subsequent question generating ability; experts and novices value different properties when judging the quality of questions; whilst investigative relevance and the type of knowledge probed may be more indicative of quality than unpredictability. However, the overarching conclusion to be drawn from the studies described here and in Chapter 4 is that question generation and, the factors that contribute towards it, is complex and a somewhat abstract process that requires substantially more investigation. At present the findings are noisy and tentative at best. In order to establish more concrete theories, the next logical step is to test the 3-dimensional model against a set of genuine forensic interview questions with clear outcomes.

Chapter 6: Applying the 3-dimensional model of question quality to real-world investigative interview questions

Abstract

The 3-dimensional model of question quality offers a potentially useful tool for judging the quality and efficacy of investigative interview questions. However, it has yet to be tested on a on a sample of real-world investigative questions with genuine interview outcomes. Transcripts were taken from interviews conducted during Ormerod and Dando's (2015) aviation security field study. Half were transcripts of successful interviewers, where a mock airline passenger had been identified, and half were transcripts of unsuccessful interviews, where a mock passenger had not been identified. The questions in the transcripts were rated by two experts using the 3-dimensional model. Additionally, they were rated for general quality. The results indicated that the ratings made using the 3-dimensional model were reliable overall and correlated strongly with the general quality ratings. Moreover, questions taken from successfully interviews were rated higher on the 3-dimensional model than questions taken from the unsuccessful interviews. This suggests the 3-dimensional model is a reliable method of rating question quality and can be used to predict the outcome of real-world investigative interviews. Future research focusing on forensic interview techniques should consider the extent to which questions are relevant, unpredictable and probing episodic knowledge. Furthermore, these findings might be used to inform emerging technological advances in the forensic field.

Introduction

The Controlled Cognitive Engagement technique (CCE; Ormerod & Dando, 2015) for investigative interviewing offers an effective method for detecting deceit. In a field study, using real aviation security agents, the technique was shown to be 24 times more effective in detecting deceptive passengers than the security methods currently in place. Whilst this is an encouraging step for the forensic community, the technique promotes the use of unexpected tests of expected knowledge which, whilst effective, are potentially challenging for an interviewer to master. To date, there is no published research investigating the factors and conditions required to generate such questions, or which components contribute towards a good quality test question. In Chapter 4, we explored the effects of topic and temporal perspective of interviews on the quality of subsequently generated questions. The results indicated that there was an interaction between topic and temporality, which was explained in terms of the scope of episodic information available to the question generator. In Chapter 5, a card sort technique was applied to a set of novice-generated test questions. This led to the development of a new 3-dimensional model for rating the quality of investigative questions, whereby they were rated for Relevance, Unpredictability, and the type of Knowledge required to provide an answer. In the present study, we take these components and apply them to pre-existing questions taken from Ormerod & Dando's field study, with the purpose of testing the 3-dimensional model of question quality, as well as the effects of scope, on a set of real-world interview questions with genuine outcomes.

The CCE technique (Ormerod & Dando, 2015), as applied to aviation security, involves a short, informal conversation between the interviewer and the interviewee. There are three stages: building rapport, information gathering, and veracity testing. In the rapport building phase, the interviewer will ask neutral questions that the

interviewee will be able to answer honestly, regardless of whether they are planning to be deceptive during the interview. This phase only occurs once in the interview and establishes a baseline for that interviewee; it allows the interviewer to observe the interviewee's natural verbal behaviour, when not under pressure.

In the information gathering phase, the interviewer will ask open questions and allow the interviewee to provide a free account. This stage may involve the interviewer prompting the interviewee, typically with more focussed questions, to provide further specific information. Question styles that allow the interviewee to provide a rich verbal account have been shown to be effective in determining veracity (Oxburgh & Dando, 2011). Finally, during the veracity testing phase, the interviewer will take the information provided in the previous phase and attempt to generate one or more 'test' questions. The test questions should seek information that it would be reasonable to expect the interviewee to know, if the account they have just provided is true. As such, these questions are referred to as tests of expected knowledge. The investigative value of these test questions is supported by Blair and colleagues' (2010) thorough investigation into the effects of context and prior knowledge on interview outcomes. Furthermore, such questions will be more effective if they are unanticipated by the interviewee (Vrij et al., 2009). The interviewer is encouraged to cycle through the information gathering and veracity testing phases several times, discussing different topics, from different temporal perspectives, in each cycle. The technique relies on detecting changes in the interviewee's verbal behaviour, especially when faced with the challenge of answering a test question, such as exhibiting gaps in expected knowledge or a reduction in response length.

In the field study conducted by Ormerod and Dando (2015), 204 mock passengers were sent through genuine airline security screening over the course of 18

months. Half were processed by security agents using the currently employed method, referred to as ‘Suspicious Signs’, and half were processed by agents that had been trained to use the CCE technique. Mock passengers were provided with a valid ticket, a flight itinerary and a basic cover story that they were asked to stick to during security screening. They were encouraged to conduct prior research regarding their cover story in order to be better prepared for the interview. The mock passengers were told that they would receive an additional £60 if they were able to successfully pass through the security screening with their deception going undetected. When agents were unsatisfied with the interview outcome, the passengers were sent for further screening (the standard process throughout aviation security). When this occurred, the passenger was recorded as having been detected; when it did not occur, the passenger was recorded as being undetected.

The results showed that the agents using the CCE technique detected 72% of mock passengers; 24 times as many as the agents using the Suspicious Signs technique. Using a matched sample of genuine passengers, it was also demonstrated that the CCE technique resulted in no more potential false alarms (genuine passengers being identified as being deceptive) than the Suspicious Signs method. Furthermore, using the CCE method resulted in passengers using more words and providing more information items, and agents using fewer words but covering more temporal perspectives, than when the Suspicious Signs method was used. Moreover, when looking at responses to the CCE interviews, there were clear differences in verbal behaviour. For example, there was no difference between the number of words used by genuine and mock passengers in response to the first open question. However, mock passengers used fewer words than genuine passengers in response to the last open question. Taken together, the

results indicated that the CCE method is highly successful in detecting deception and a potentially invaluable tool for forensic settings.

Whilst the evidence gained by Ormerod and Dando (2015) points to the efficacy of the CCE technique, the test question phase is arguably a difficult and fundamentally creative skill to learn. In the aviation security setting, it requires an agent to be able to carefully listen to the interviewee's account on a given topic, determine some key details within that account and then generate an unexpected question that it would be reasonable to expect the interviewee to know, based on the information they have provided. Moreover, they need to be able to do this several times, across various topics and temporal perspectives, for each interview. The findings of the study indicate that this is certainly possible with the right training. Additionally, in Chapter 4, novice participants who had been shown a 10-minute training video that covered the basics of the CCE technique were shown to generate better quality test questions than participants who had not watched the video. However, what is currently unclear is the extent to which there was variation, in terms of the quality of test questions, within the CCE interviews conducted in the study. Naturally, one would expect variation in quality. However, the important factor is whether this variation affects interview outcomes. If this were to be the case, it will be vitally important to establish the dimensions of question quality that contribute towards positive interview outcomes, in order to better equip those who are training or using the CCE technique.

In chapter 5, the first exploratory steps were taken in terms of establishing such components of question quality. Using a card sort to categorise a large sample of novice-generated test questions, seven distinct question types were identified. The authors determined three dimensions that, when a low, medium, or high ranking was assigned to the seven question types, were adequately able to capture the differences

between them. The three dimensions were Relevance, Unpredictability, and type of Knowledge probed. Police in the UK are trained to keep interviews relevant to the investigation in hand. Taking the other two dimensions into account, if an interviewer were to ask a suspect "how long does it take to drive from your house to the nearest supermarket?" that might be entirely unpredictable and certainly requires an episodic knowledge of the area. However, if that information is irrelevant to the investigation, the answers it yields may have little value to the interviewer. As such, relevance is likely to be an important dimension of question quality. Unpredictability is valuable because it removes the option for a deceptive interviewee to rely on a prepared script. Given the opportunity, a liar will instinctively attempt to anticipate potential questions and prepare plausible sounding responses to them (Vrij et al., 2017). If they are prevented from applying this tactic, the interviewee is forced to lie spontaneously, which is a more cognitively demanding task and has been shown to elicit changes in verbal behaviour (Porter & Yuille, 1996; Vrij, 2014). Finally, asking questions that probe episodic knowledge, as opposed to general semantic knowledge, also presents a deceptive interviewee with a challenge. Given that lies are endogenously conceived, they are more likely to contain general semantic information, since the deceiver does not have unique episodic information available (Oberlader et al., 2016; Sporer, 2016). Therefore, lines of questioning that require the interviewee to have genuinely experienced the event in question should be difficult for a deceptive interviewee to answer and, as such, might lead to changes in verbal behaviour (Sporer, 2016).

In Chapter 5, questions generated by novices and experts were rated by four interviewing experts using this 3-dimensional model. The results were mixed. In terms of reliability, there was medium strength agreement between the four raters for Relevance and Unpredictability. However, both fell slightly below the pre-determined

threshold set for the study. Despite this, there was a strong agreement found between the raters for the Knowledge dimension, as well as for the scale overall with the three dimensions summed. The judges also provided a rating for general quality and the results showed that three of the judges' quality ratings correlated well with their own ratings on the overall 3-dimesional scale, though one rater's scores did not correlate at all. In terms of the individual dimensions, Unpredictability did not correlate with general quality for any of the four raters, whilst Relevance and Knowledge appeared to contribute more strongly.

Taken together, this provided mixed support for the model, generating concerns regarding its reliability and association with quality. However, one potential issue with the methodology was that the three dimensions were only explained via a set of brief instructions, which may have contributed towards the reliability and applicability issues. Such issues might be overcome with more thoroughly explained definitions. Another issue is that there was no real outcome measure by which to compare the ratings. As such, it was not possible to determine whether questions that were rated highly across the three dimensions led to objective, positive interview outcomes. In order to judge the value of the 3-dimensional model, there is a need to establish, not only whether it can be used to discern between good- and poor-quality questions, but also whether it is capable of predicting the success of an interview outcome. If this were the case, it would provide the investigative community with a useful tool for judging the quality of test questions, as well as assessing the likelihood that certain questions will result in a positive outcome.

Chapter 5 also investigated the effect that the scope of episodic information, inherently available to the question generator, had on the quality of the questions that were generated. This was based on the findings of Chapter 4, whereby it was

demonstrated that there was an interaction effect between the topic of interview and the temporality of the topic on question quality. Participants were shown a series of information gathering interview clips whereby the interviewees discussed four topics (home, hobbies, travel and work), from either a past or a present perspective, and were asked to generate a test question after each. The results indicated that, for interviews where the topic was travel, questions were rated higher in quality when it was discussed from a past temporal perspective compared to a present perspective. Conversely, for the work interviews, a present perspective was found to be preferable. This was explained in terms of the scope of episodic information available within those topic/temporality combinations. For example, with travel, asking about current travel plans limits the available episodic knowledge to the planning of the event, given that the trip has not yet taken place. In this case, the scope of episodic information available would be narrow. In contrast, discussing travel that occurred in the past does not have this restriction; the individual has already experienced that trip and there should be a wealth of episodic knowledge available to question them on. In this case, therefore, the scope is broad.

This theory was tested in Chapter 5. Scope was manipulated so that a group of participants carried out four tasks, either with specific instructions (narrow scope) or more general instructions (broad scope). The participants were subsequently interviewed about the tasks. Clips of these interviews were shown to a separate group of participants who were required to generate a test question after each. Following this, four expert judges rated the questions according to the 3-dimensional model. The results indicated that there was an effect of scope on the ratings. However, this went against prediction; tasks where the scope was narrow led to higher rated questions than tasks when the scope was broad. Moreover, looking at the individual dimensions of the model revealed that this effect was mainly due to a difference on the Knowledge scale, with

questions generated for the narrow scope tasks being judged as probing more episodic information than questions generated for the broad tasks. This contradictory finding was explained in terms of the context of the interview. A broad scope appeared to be beneficial when the interviews were informal discussions about general topics. However, when the interviews were more formal in context and referred to a specific incident, a narrow scope perhaps provided a more constrained framework with which to probe specific episodic details.

In the present study, transcripts were taken from the security screening interviews conducted in Ormerod and Dando's (2015) field study. Half of these transcripts were taken from interviews where the mock passenger was identified, and half were taken from interviews where the mock passenger was not identified. This not only provided a set of investigative questions with a more applicable, professional context than in previous chapters, but also provided a genuine outcome measure with which to determine whether higher rated questions led to more positive interview outcomes. Furthermore, the questions were coded for topic and temporality so that it was possible to further examine the effects of scope on question quality. The transcripts were rated by two interviewing experts, who were blind to condition, for both general quality and using the 3-dimensional model. The purpose was to assess the 3-dimensional model in three key ways: firstly, to establish whether the three dimensions can be rated reliably when judges are provided with clear, explicit definitions; secondly, to determine whether ratings on the scale correlate with the ratings on the more implicitly defined measure of general quality; and finally, to determine whether questions taken from interviews with a successful outcome were rated higher than those taken from unsuccessful interviews. Additionally, the study was designed to further

investigate the effects of topic and temporality on question quality when the interviews are informal but forensically motivated.

This led to seven hypotheses. It was predicted that there would be an acceptable level of inter-rater reliability found between the two judges' ratings of general quality and the ratings given on the 3-dimension model overall (Hypothesis 1). It was also predicted that there would be an acceptable level of inter-rater agreement for the ratings of the three individual dimensions (Hypothesis 2). Given that the three-dimensional model should be measuring the quality of the questions, there will be a strong positive correlation found between the ratings on the 3-dimensional scale overall and the general quality ratings (Hypothesis 3), as well as a positive correlation between ratings of general quality and the three individual dimensions (Hypothesis 4). If both scales are adequately judging question quality, it would be reasonable to expect that questions which come from transcripts in which the mock passenger was detected would have a higher general quality rating, as well as a higher rating on the 3-dimensional model overall, than questions taken from transcripts in which the mock passenger was not detected (Hypothesis 5). Furthermore, this effect should also be evident within the three individual dimensions, if each dimension does contribute towards question quality (Hypothesis 6). Finally, based on the findings of Chapter 4, we expect to find an interaction effect between interview topic and temporality of topic on the quality of the questions (Hypothesis 7).

Method

Participants

Two male interviewing experts were recruited to rate the questions contained in the transcripts. Rater 1 was aged 33 and had 4 years interviewing experience. Rater 2

was 56 and had 6 years interviewing experience. Both were blind to condition and took part voluntarily.

Materials

The questions rated in the present study were taken from transcripts of interviews conducted in Ormerod and Dando's (2015) aviation security study. The study tested the use of the CCE interview technique in a real-world airport situation by sending 204 mock flight passengers through a genuine security screening process. The mock passengers were either screened according to the current practise for the airline (Suspicious Signs screening) or according to the CCE method. The main outcome measure was whether the security agent identified the deception (i.e. whether the passenger was sent for further screening). Each interview was recorded and subsequently transcribed as part of the original study. Transcripts were presented in table format using Microsoft Word software, with the interviewer's questions in one column and the interviewee's response in the adjacent column. For the present study, in each transcript, each new information gathering section was highlighted in yellow, with the following related test questions phase highlighted in red. Each set of questions was then coded for topic and temporality of topic.

In total, 80 transcripts taken from the CCE interviews were selected to be utilised in the present study. Of these, 40 were from interviews where the deception was identified, and 40 were from interviews in which the deception was not identified. Each transcript was assigned a number in a random order and any identifying information regarding deception identification was removed. Whilst audio recordings of the interviews may have been preferable, this was not possible. Due to data protection rules, and the fact that the audio recordings were mainly collected in 2012, they had already

been deleted to protect participant anonymity, as specified by the ethical clearance obtained by Ormerod and Dando (2015). While using transcripts raises the possibility of errors, the interviews used here were short (typically 2-3 minutes). As such, we are confident that the risk of errors was minimal, when compared with longer police suspect interviews.

Design

A mixed design was employed for the ratings. Both judges rated the questions in all 80 of the transcripts, for both general quality as well as using the 3-dimensional model developed in Chapter 5. Transcripts were rated in a random order, with half being taken from screening interviews where the mock passenger was identified (n = 40) and half taken from screening interviews where the mock passenger was not identified (n = 40).

Procedure

The judges first attended a training exercise, in which they were given clear, objective definitions of the three dimensions used in the 3-dimensional model. They were provided with example questions and then discussed ratings across the dimensions for each question. Discussions and examples continued until the researchers were satisfied that the judges correctly understood each dimension. This process lasted around 90 minutes in total. Following this, they were each provided with a computer folder containing all 80 of the transcripts, each assigned a number in random order. They were also sent a Microsoft Excel file which listed each transcript number, followed by a row for each of the question sets within that transcript, in the order with which they appeared in the document. Next to each question set were three columns in which they were asked to provide a rating, on a 1-7 scale, for Relevance (1 = not at all

relevant; 7 = very relevant), Unpredictability (1 = very predictable; 7 = not at all predictable) and Knowledge probed (1 = entirely semantic knowledge probed; 7 = entirely episodic knowledge probed), as outlined in Chapter 5. They were instructed to open each transcript in turn, read the first information gathering section highlighted in yellow, then the subsequent test questions phase highlighted in red, and then provide ratings for that set of test questions as a whole in the Excel sheet provided. They continued to do this until each information gathering/test questions set in the transcript had been rated. This was repeated until all 80 transcripts had been completed. Subsequently, the same procedure was followed for the general quality ratings. This was completed one week later in order to reduce the chance of the judge's quality ratings being influenced by their ratings on the 3-dimensional model. Additionally, to control for order effects, the transcripts appeared in a new random order.

Data Analysis

Inter-rater reliability was examined using the same procedure outlined in Chapter 4, using Intraclass Correlation Coefficients (ICC). For each set of ratings, a two-way random-effect model based on average ratings and consistency (ICC(C,k)) was applied across the entire set of ratings (i.e., ratings for every question set in each transcript were included in this analysis). Following the guidelines presented by Cicchetti and Sparrow (1990), reliability was considered to be acceptable when correlations were above a .70 threshold.

Following reliability analysis, the two judges' scores were averaged where appropriate. Subsequently, analysis was conducted in two ways: firstly, via looking at the average rating across all question sets for each transcript and secondly by only including the question set with the highest average general quality score for each

transcript. This second analysis was performed because, despite there being no significant difference between the average number of question sets included in the Detected (M = 2.58, SD = 0.96) and Non-Detected transcripts (M = 2.23, SD = .70; $t(78) = 1.87$, $p = .07$), there was still the potential for this factor to bias the average ratings, especially when a good quality question set appeared in a transcript with several other poor question sets.

The final analysis assessed the effect of topic and temporality on the quality of the questions. This analysis was performed using the quality rating for each set of questions in each transcript. Given that the data was taken from a pre-existing field study, it was not possible to control for the number of question sets falling into each category. For topic, there were six categories: Travel Plans ($n = 59$), Employment ($n = 55$), Family and Friends ($n = 27$), Hobbies and Interests ($n = 9$), Education ($n = 22$) and Hometown ($n = 11$). For temporality, 35 question sets fell into the past category, whilst 148 fell into the present category. Nine additional transcripts were excluded from this analysis as they did not fall into any of the above categories.

Results

Inter-Rater Reliability

General quality. To determine whether there was an acceptable level of reliability between the judges' ratings of general quality, an ICC(C,k) analysis was conducted. The ICC was .82, 95% CI [.76, .86], which indicates good reliability and meets the acceptable threshold. As such, the mean of the two judges' scores was calculated for each rating and this average measure was used in subsequent analysis.

3-dimensional scale. To examine the level of inter-rater agreement for the 3-dimension rating scheme, a series of ICC(C,k) analyses were conducted, first for the

individual dimensions and finally for the mean rating of the three dimensions for each question set. The ICC(C,k) for Relevance was .25, 95% CI [.01, .28], which indicates poor reliability and does not meet the acceptable threshold. The ICC(C,k) for Predictability and Knowledge were .84, 95% CI [.78, .88] and .80, 95% CI [.71, .83], respectively. Both indicate good reliability and met the acceptable threshold. Finally, taking the average score of the 3-dimensional model, the ICC(C,k) was .84, 95% CI [.78, .88], again indicating good reliability and meeting the threshold.

Despite the finding that ratings of Relevance showed poor reliability, the average ratings, using the scale as a whole, were found to be reliable. Therefore, the two judges' ratings were combined to form a mean score for each dimension, and in turn, an average overall mean of the scale was calculated using these combined scores. All subsequent analysis was conducted using these average scores.

Transcript Average Ratings

Correlation between scales. To examine the correlation between the average general quality ratings and ratings made using the 3-dimensional model, a series of Pearson Correlation Coefficient analyses were conducted. There was a strong positive correlation between the average general quality ratings and the average ratings given on the 3-dimensional scale overall, $r = .92$, $n = 80$, $p < .001$. Looking at the individual components, there was also a strong positive correlation between the general quality ratings and the predictability ($r = .88$, $n = 80$, $p < .001$) and knowledge ratings ($r = .93$, $n = 80$, $p < .001$). For relevance, there was a small-to-medium positive correlation, $r = .47$, $n = 80$, $p < .001$.

Deception Detection. In order to examine differences in the ratings of general quality between transcripts where the deception was detected and those where the

deception went undetected, an independent t-test was conducted on the average rating for each transcript. For interviews where the mock passenger was detected, questions were rated as significantly higher in quality (M = 4.37, SD = 1.08) than for interviews where the mock passenger was not detected (M = 3.35, SD = 0.83), $t(78) = 4.75$, $p < .001$, $d = 1.06$, 95 % CI [0.58, 1.52].

A further independent t-test was conducted to examine the difference between ratings on the 3-dimension scale. The analysis again revealed that, for interviews where the mock passenger was detected (M = 4.37, SD = 0.86), the questions were rated significantly higher on the 3-dimensional scale than when the mock passenger was not detected (M = 3.42, SD = 0.68), $t(78) = 5.49$, $p < .001$, $d = 1.23$, 95% CI [0.74, 1.69].

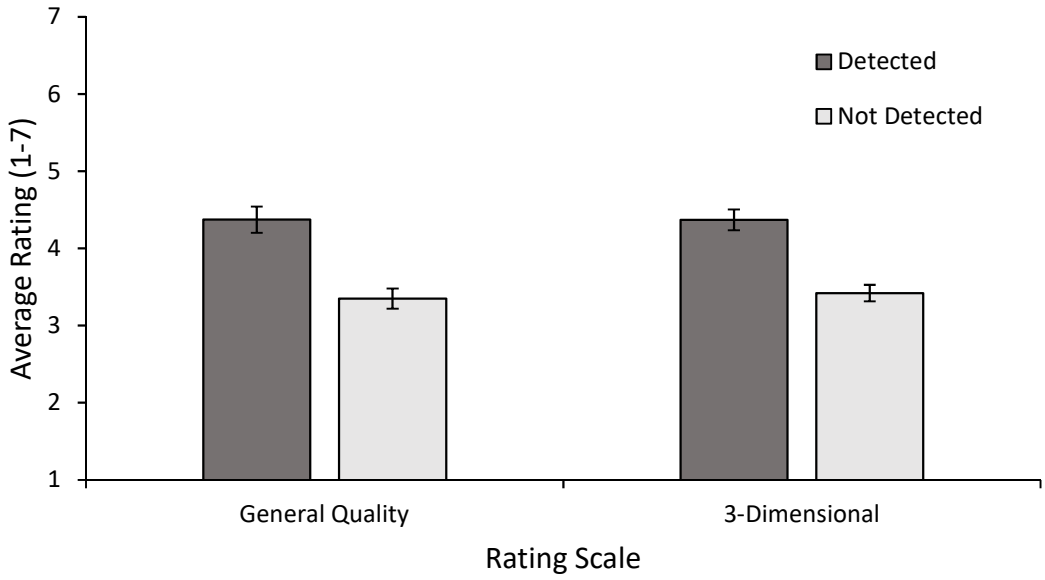

Figure 1. Average ratings for questions in deception detected and deception not-detected transcripts, on both the general quality scale and the 3-dimensional model (transcript average ratings). Error bars represent +/- 1 SEM.

To examine group differences on the three individual components of the model, a series of independent t-tests were conducted. There was a significant difference

between Relevance ratings, with questions from transcripts where the mock passenger was detected (M = 4.84, SD = 0.42) being rated as more relevant than questions from transcripts where the mock passenger was not detected (M = 4.48, SD = 0.51), $t(78) = 3.45$, $p = .001$, $d = 0.77$, 95% CI [0.31, 1.22] (though this finding should be treated with caution given the lack of inter-rater reliability). There was a significant effect on ratings of Unpredictability, with questions on the detected transcripts (M = 4.00, SD = 1.24) being rated as more unpredictable than questions on the non-detected transcripts (M = 2.75, SD = 0.93), $t(78) = 5.11$, $p < .001$, $d = 1.14$, 95% CI [0.66, 1.60]. Finally, there was also a significant effect on the Knowledge ratings, with questions on the detected transcripts (M = 4.27, SD = 1.17) being rated as probing more episodic information than questions on the non-detected transcripts (M = 3.03, SD = 1.04), $t(78) = 4.99$, $p < .001$, $d = 1.12$, 95% CI [0.64, 1.58].

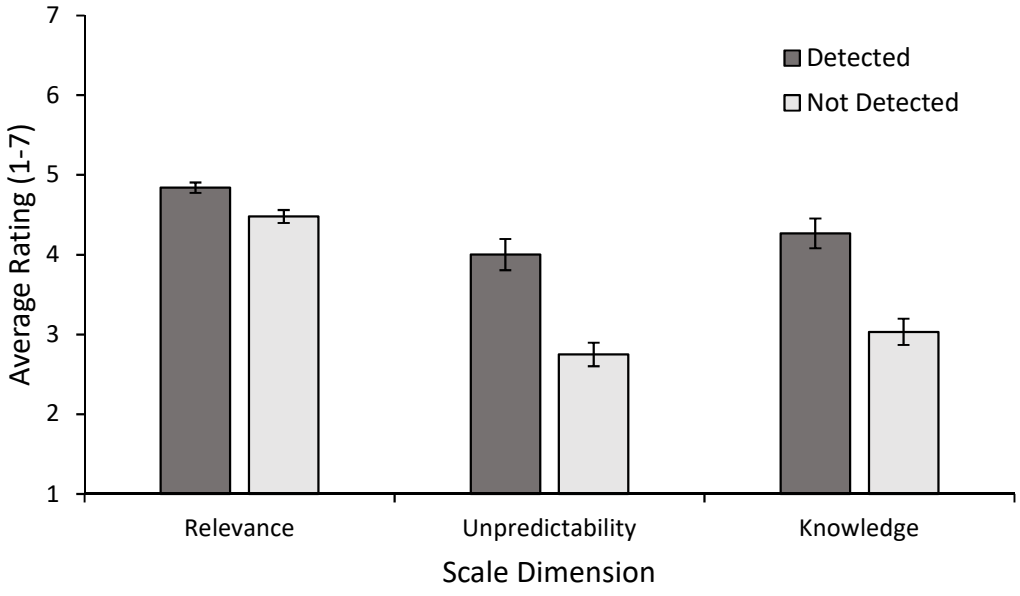

Figure 2. Average ratings for questions in deception detected and deception not-detected transcripts, on each component of the 3-dimensional model (transcript average ratings). Error bars represent +/- 1 SEM

Best Question in Set

The same set of analyses described above were repeated using data whereby the highest rated question set in each transcript was taken, as opposed to the transcript average.

Correlation between scales. To examine the correlation between the average general quality ratings and ratings made using the 3-dimensional model, a series of Pearson Correlation Coefficient analyses were conducted. A strong positive correlation was found between the average general quality ratings and the average ratings given on the 3-dimensional scale overall, $r = .91$, $n = 80$, $p < .001$. Looking at the individual components, there was a small positive correlation between general quality ratings and the relevance ratings, $r = .35$, $n = 80$, $p < .001$. There was a strong positive correlation between the general quality ratings and the predictability ratings ($r = .87$, $n = 80$, $p < .001$). Likewise, there was also a strong positive correlation with the knowledge ratings ($r = .87$, $n = 80$, $p < .001$).

Deception detection. An independent t-test was conducted to examine the difference in general quality ratings between questions in detected and non-detected transcripts. For interviews where the mock passenger was detected, questions were rated as significantly higher in quality (M = 5.16, SD = 1.04) than for interviews where the mock passenger was not detected (M = 3.81, SD = 0.92), $t(78) = 6.14$, $p < .001$, $d = 1.37$, 95 % CI [0.88, 1.85].

A further independent t-test was conducted to examine the difference between ratings using the 3-dimensional model. In interviews where the mock passenger was detected (M = 4.96, SD = 0.92), the questions were rated significantly higher on the 3-

dimensional model overall than when the questions where the mock passenger was not detected (M = 3.83, SD = 0.79), $t(78) = 5.84$, $p < .001$, $d = 1.32$, 95% CI [0.82, 1.79].

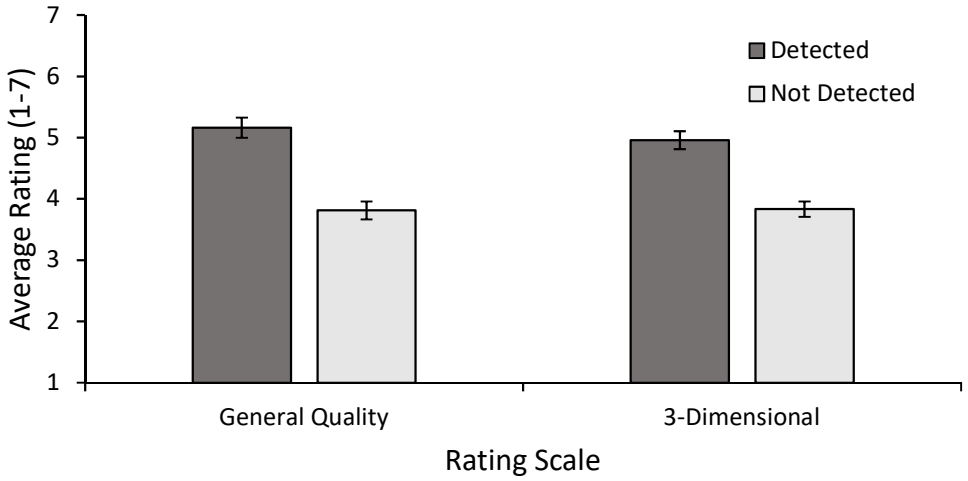

Figure 3. Average ratings for questions in deception detected and deception not-detected transcripts, on both the general quality scale and the 3-dimensional model (highest question set per transcript). Error bars represent +/- 1 SEM.

Looking at the three individual components of the model, there was a significant difference between Relevance ratings, with questions from deception detected transcripts (M = 5.01, SD = 0.49) being rated as significantly more relevant than questions from deception not-detected transcripts (M = 4.78, SD = 0.54), $t(78) = 32.06$, $p = .04$, $d = 0.45$, 95% CI [-0.002, 0.89] (this finding should be treated with caution given the lack of inter-rater reliability). There was a significant effect on the ratings of Unpredictability, with questions on the detected transcripts (M = 4.74, SD = 1.44) being rated as more unpredictable than questions on the non-detected transcripts (M = 3.26, SD = 1.28), $t(78) = 4.85$, $p < .001$, $d = 1.09$, 95% CI [0.61, 1.54]. Finally, there was also a significant effect on Knowledge ratings, with questions on the detected transcripts (M = 5.13, SD = 1.21) being rated as probing more episodic information than questions on the non-detected transcripts (M = 3.46, SD = 1.17), $t(78) = 6.25$, $p < .001$, $d = 1.40$, 95% CI [0.90, 1.88].

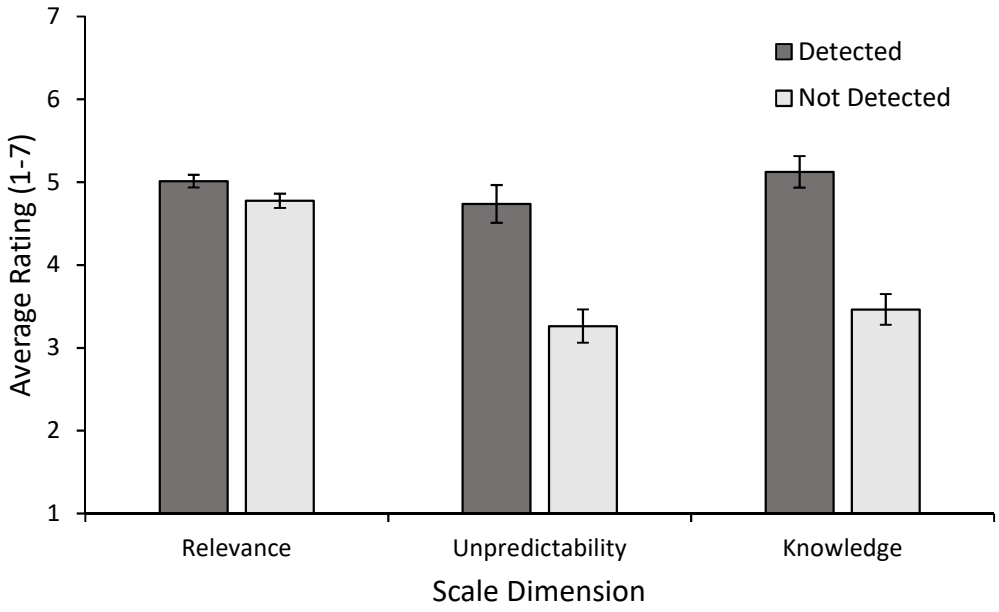

Figure 4. Average ratings for questions in deception detected and deception not-detected transcripts, on each component of the 3-dimensional model (highest question set per transcript). Error bars represent +/- 1 SEM

Topic and Temporality

In order to determine whether the topic of the question sets and the temporality of the topic had any effect on the general quality ratings, a 6 (Topic: Travel Plans, Employment, Family and Friends, Hobbies and Interests, Education, and Home Town) × 2 (Temporality: Past vs Present) ANOVA was conducted. The results revealed a significant main effect of Topic, $F(5, 171) = 4.48$, $p < .001$, $\eta_p^2 = .79$. However, there was no significant main effect of Temporality, $F(1, 171) = 0.25$, $p = .62$, nor was there a significant interaction, $F(5, 171) = 1.38$, $p = .24$.

To further examine the effect of topic, a Tukey HSD post-hoc test was conducted (see Figure 5). The analysis revealed that when the topic of the question set was Travel Plans (M = 3.32, SD 0.89), the quality of the questions were rated significantly lower than when the topic was Employment (M = 4.33, SD = 1.18), and

when it was Hobbies and Interests (M = 5.11, SD = 1.64), both at $p < .001$. Additionally, there was a marginal difference in ratings between Travel Plans and Education (M = 4.16, SD = 1.31; p = .05). Ratings when the topic was Family and Friends (M = 3.35, SD = 1.17) were also significantly lower than when the topic was Employment ($p = .01$), as well as when the topic was Hobbies and Interests ($p = .002$). All other comparisons were not significant. This set of analyses were repeated using the average ratings taken from the 3-dimenionsal scale, resulting in the same pattern of findings.

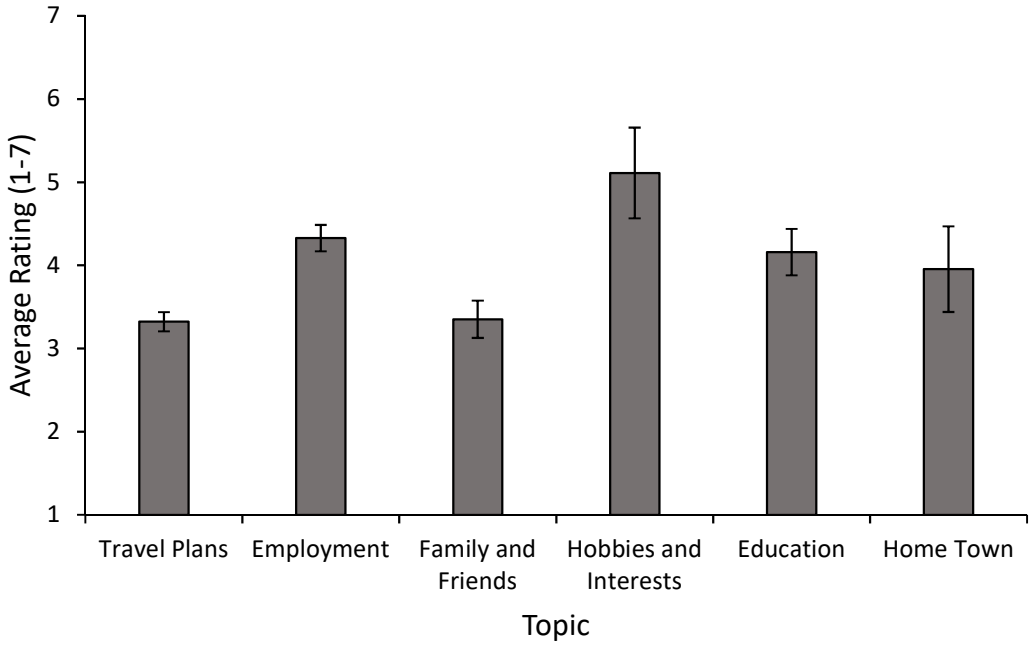

Figure 5. Average general quality rating across topic of question set. Error bars represent +/- 1 SEM.

Discussion

The present study applied the 3-dimensional model, designed in Chapter 5, to a set of aviation security screening interview questions with genuine outcomes. This was performed in order to test the reliability of the scale and to determine its efficacy in predicting interview outcome success. Questions taken from the transcripts of 40

successful and 40 non-successful CCE security screening interviews, conducted in Ormerod and Dando's (2015) field study, were rated by two experts using both the 3-dimensional model and the more subjective general quality scale. Additionally, the questions were coded for topic and temporal perspective in order to examine the effect that these factors have on question quality, especially in interaction. A system that is able to accurately judge the quality and efficacy of investigative questions, across three clearly defined dimensions, would be valuable tool for the investigative community, in terms of improving training procedures or in relation to the planning stage of suspect interviews.

The results indicated an improvement in reliability for the 3-dimensional model, compared to the findings of Chapter 5. There was a strong, positive correlation between ratings using the 3-dimensional model and the ratings of general quality, which increases confidence that the 3-dimensional model is effectively capturing differences between good- and poor-quality questions, and doing so in a more practical, objective manner than the general quality scale. Moreover, the findings revealed that the 3-dimensional model was sensitive to objective differences in quality; questions taken from interviews where the mock passenger has been successfully identified were rated higher on the model overall, as well as for each individual dimension, than questions taken from unsuccessful interviews. Finally, the present study failed to find an interaction between topic and temporal perspective on question quality. This finding can potentially be accounted for by the fact that it was not possible to control these variables. As such, there was a large bias towards questions with a present temporal perspective, which may have negated any potential interaction.

In terms of the reliability of the 3-dimensional model, strong inter-rater agreement was observed between the combined scores of the three dimensions.

Likewise, there was strong agreement for ratings on the general quality scale. Taken together, this supports the first hypothesis and increases confidence in the reliability of both scales. Additionally, there was a high level of agreement found between the ratings given for both the Unpredictability and Knowledge dimensions. However, the ratings for Relevance showed little agreement. Whilst this is a concern in terms of the reliability of this dimension, it can perhaps be explained as an effect of the interview context. In aviation security screening, almost any topic that might raise challenges concerning the presented identity or intent of the interviewee will be of relevance. As such, Relevance in this context is perhaps less discriminatory than the other two dimensions. Looking at the average ratings on each dimension, Relevance had far less variation among scores, exhibited by a substantially lower standard deviation. Therefore, small disagreements between the two raters in this context become more pronounced, leading in turn to a more sensitive measure of inter-rater reliability. Overall, this provides mixed support for Hypothesis 2. It seems necessary to further test the model across alternative investigative interview contexts (e.g., police interviews with persons of interest) in order to fully examine the value and contribution of each dimension, as the objectives and context of the interview change.

There was a strong correlation between the ratings on the 3-dimesional model as a whole and the general quality ratings. This finding was observed using both the transcript average ratings and the highest rated question set in each transcript. This supports Hypothesis 3 and further increases confidence in the model's efficacy in capturing the quality of investigative questions. Moreover, there was a strong positive correlation found between the general quality ratings and both the Unpredictability and Knowledge ratings. There was also a small-to-medium positive correlation found with the Relevance ratings. Overall, this supports Hypothesis 4 and again indicates that the 3-

dimensional model, as both an overall measure and as three separate dimensions, can capture differences between good- and poor-quality questions. Previous chapters have shown that the general quality measure can be both reliable and effective in determining certain factors that contribute towards question quality. However, its practical use is limited; the scale is subjective in nature, open to each judges' interpretation of quality, and does not delineate between the various facets that contribute towards question quality. For example, if looking at scores on the general quality scale, it would not be possible to determine why a judge had given a particular question a low rating, nor a high rating. In contrast, with the 3-dimensional model, it is possible to gain insight into the decision-making process which, in turn, allows for further exploration into the factors associated with good quality questioning. As such, having a more objective, finely-tuned and practical method by which to judge question quality is a positive step in determining the creative factors associated with question generation.

The main investigation in the present study was to establish whether there was variation in terms of the quality of questions asked in the CCE interviews, and whether such variation affects the outcome of the interview. The results revealed that questions taken from transcripts where the mock passenger was identified were rated higher for both general quality and on the 3-dimensional model as a whole than questions that were taken from non-successful interviews. This effect was found for both the transcript average ratings and for the analyses which used the highest rated question set in each transcript, supporting Hypothesis 5. Moreover, questions from successful transcripts were rated higher in each of the three dimensions individually than questions from the non-successfully transcripts, supporting Hypothesis 6. Taken together, these findings have important implications for the CCE technique.

Ormerod and Dando (2015) were able to demonstrate that the CCE technique was substantially more effective in detecting deception in aviation security screening than the currently employed Suspicious Signs technique. The results demonstrated that CCE questions led to increased differences in both the interviewers' and interviewees' verbal behaviour, which in turn emphasised differences between genuine and mock passengers' verbal behaviour. However, the study did not investigate whether there was variation in terms of the quality of questions asked within the CCE interviews themselves. The results presented here suggest that there is a detectable variation in quality and, furthermore, this variation can affect the outcome of the interview. Questions from interviews with a successful outcome were more likely to be relevant, unpredictable and probing episodic knowledge. Therefore, there is a need for CCE training courses to reflect these findings. Those learning the technique and hoping to apply it to forensic situations should be taught to consider these three dimensions when generating test questions. Furthermore, the three dimensions are not solely applicable to CCE interviews. In the UK, police officers are routinely trained in the PEACE model of investigative interviewing. The first step in this model is Planning and Preparation, which has been shown to improve interview quality and lead to more positive outcomes (Walsh & Bull, 2010). Interviewers are encouraged to make a written interview plan which includes the range of topics to be covered, the points required to prove the potential offence, and information which may assist the enquiry (Authorised Professional Practice, 2019). Arguably, based on the findings presented here, it may benefit the interviewer to also consider how to phrase key questions in a way that is relevant, unpredictable and probing episodic knowledge.

In Chapter 4, we found that the there was an interaction between the topic of interview and the temporal perspective of the topic on the quality of questions that had

been generated from CCE-style information gathering interviews. This was explained in terms of the episodic scope inherently associated with certain topic/temporality combinations; discussions with a broad episodic scope provide the interviewer with a wider range of episodic information points to draw on when generating test questions. In Chapter 5, which utilised interviews of a more forensic nature regarding a specific incident, we found that scope did affect the quality of question generation, though in the opposite direction than predicted. The narrow scope interviews resulted in higher quality questions being generated than the broad scope interviews. This suggests that scope is a factor that can affect the quality of question generation. However, whether a narrow or broad scope is beneficial may depend on the context of the interview. We theorised that a broad scope may improve the quality of questions during informal discussions of broad topics, but a narrow scope may be required when discussing a specific incident.

The present study failed to find support for the effect of scope. There was an effect of topic on the quality ratings, however, there was no effect of temporality and, more importantly, there was no interaction between the factors. In terms of the effect of topic, the results indicated that questions focussing on Employment and Hobbies and Interests led to higher rated questions than those which focussed on Travel Plans and Family and Friends. The main limitation of this analysis was that it was not possible to control for these factors in advance. The transcripts were taken from genuine aviation security screening interviews. Whilst this allowed us to conduct an invaluable analysis of investigative questioning within a professional, applied context, it came at the cost of control. As such, there was a lack of parity between the number of questions falling into each group. For example, 80% of the questions used in the sample were asked from a present temporal perspective. This may account for the lack of an interaction effect

between topic and temporality. Take, for example, the Travel-based questions, of which 92% were asked from a present perspective. In Chapter 4 we argued that asking an interviewee to describe their current travel plans restricts the available episodic information to the planning of that travel, resulting in the finding that travel questions were rated higher when asked from a past perspective. The lack of past-based travel questions in the current study may account for the low overall question rating for that topic, and more broadly, for the lack of interaction between topic and temporal perspective found overall.

Whilst the questions used in this study represents a somewhat more ecologically valid dataset than the laboratory-based studies conducted in previous chapters, aviation security screening is still substantially different in context to the investigative interviews conducted during forensic investigations. CCE involves an informal discussion surrounding general topics that are not determined in advance (Ormerod & Dando, 2015) and, as such, the findings presented here cannot be directly applied to forensic interviews involving specific, potentially illegal, incidents. The next step, in terms of testing the reliability and efficacy of the 3-dimensional model, will be to access transcripts from genuine police suspect interviews. This will allow us to determine whether the model is effective in judging the quality of more forensically motivated questions. Ideally, this would be accompanied by an outcome measure by which to judge the success of the interviews. For example, Leahy-harland and Bull (2017) examined the strategies and types of questions used by police interviewers in 56 genuine suspect interviews, with an outcome measure of whether the suspect made a full admission, a partial admission or a denial of the incident under investigation. Walsh and Bull's (2010) analysis of the techniques used during suspect interviews distinguished between 'desirable outcomes', whereby either a full confession has been

lawfully obtained or that the interview had been conducted as thoroughly as ethically possible, and 'less desirable outcomes', such as non-responsive interviewees, denials that were not thoroughly tested, and partial admissions which failed to be fully established. By adopting a similar method, it would be possible to determine whether questions that rate highly on the 3-dimensional model lead to positive outcomes in a forensic context.

An interesting potential implication of this research relates to technological advances on the horizon involving Natural Language Processing (NLP). There is work currently being undertaken at the University of Sussex to develop an NLP-based application that is capable of listening in to an interview, processing details within an interviewee's statement and then scanning the internet to retrieve useful information regarding those details in real time. Similar applications have been developed to allow people to report crimes anonymously online (Chih, Iriberri, & Leroy, 2008). This could be applied to a technique such as CCE and be used to help the interviewer generate questions with which to challenge an interviewee's account. However, for a system such as this to be effective, it would require a method by which to rank the gathered information so that the interviewer is only provided with useful details. The 3-dimensional model, tested here, could potentially form the basis of this ranking.

The present study was designed to further investigate the reliability and efficacy of the 3-dimensional model developed in Chapter 5. The results regarding the reliability of the model were mixed; there was strong agreement between the judges in terms of the overall model ratings, as well as for the Unpredictability and Knowledge dimensions. However, the context of the CCE-style interviews resulted in a lack of agreement on the Relevance dimension. Despite this, there was a strong correlation found between ratings using the 3-dimensional model and the ratings of general quality.

Furthermore, ratings made using the model were effective in distinguishing between successful and non-successful interviews.

Taken together, these results provide tentative support for the reliability and efficacy of the model in determining the quality of investigative questions in an objective, practical way. The findings reveal important implications for those training or currently practising the CCE technique, as well as the investigative community more broadly, in that a good quality question will be relevant, unpredictable and tapping into episodic knowledge. Continued research, involving questions from genuine police interviews with real outcome measures, is still required. However, the present study represents a positive step in terms of determining the factors involved in generating good quality investigative interview questions and may even assist future technological advances in interviewing.

Chapter 7: General Discussion

The thesis presented here represents an exploratory investigation into the conditions, factors, and dimensions that contribute towards the generation of good quality investigative interview questions. In 1993 the UK police force introduced the PEACE model of interview practice which, among several other important insights, reinforced the idea that there are distinct question types, each potentially affecting the outcome of an interview. However, given that the PEACE model is a general framework for interviewing, it does not specifically focus on the content of interview questions, nor the ability to detect deception. Having techniques that allow an interviewer, or independent observers, to determine the veracity of an interviewee's account is vitally important (Gudjonsson, 2003). However, research indicates that humans tend to perform poorly (Bond & DePaulo, 2006). One reason for this is that they tend to focus on non-verbal cues to deceit (Global Deception Research Team, 2006), which have routinely been shown to be unreliable (DePaulo et al., 2003). As such, most recent research in this area has focused on verbal cues to deceit, with techniques such as Unanticipated Questions (UQ; Vrij et al., 2009), Strategic Use of Evidence (SUE; Granhag et al., 2007), and Controlled Cognitive Engagement (CCE; Ormerod & Dando, 2015) all claiming to be effective methods for eliciting verbal behaviour changes from deceptive interviewees. Whilst there is empirical support for these claims, there is little research investigating the factors that enhance or inhibit an individuals' ability to generate the types of questions suggested by these techniques.

The present thesis sought to investigate these factors. First, it explored the dimensions by which one can distinguish between a good-quality question and a poor-quality question. Secondly, it set out to determine whether reliable measures of question quality can provide ratings of interview questions. Finally, it explored the factors that

may enhance or inhibit one's ability to generate good-quality interview questions. Accurately determining the dimensions and factors that affect question generation, and developing a model by which to rate the quality of interview questions, provides the investigative community with valuable information that could be used to enhance the planning and preparation of an interview, increase the chances of obtaining a successful interview outcome, and could potentially be used to inform current technological advances designed to assist the investigative interview process. In order to arrive at this stage, three main research questions were explored:

1) *What are the dimensions by which good- and poor-quality questions differ?*
2) *Is it possible to rate those dimensions objectively?*
3) *What factors enhance or inhibit one's ability to generate good-quality questions?*

Summary of the Studies

The thesis comprised five papers, each presented as a discrete chapter, designed to explore the dimensions that separate good- and poor-quality interview questions and the factors that affect individuals' ability to generate such questions. Chapters 2 and 3 focused on one of the most well-researched potential dimensions of question quality: unexpectedness. Vrij and colleagues (2009) work on the UQ approach was initially investigated via a systematic review in Chapter 2, finding that UQs do reliably elicit verbal cues to deceit, such as differences in statement length, level of detail and plausibility. However, three issues were identified: there was no measure of interviewer veracity detection accuracy reported in any study; there was a lack of cohesion regarding the types of UQs used; and the cognitive load theory, proffered by Vrij (2014), found only mixed support.

Chapter 3 presented a two-part study investigating these issues. The findings revealed that UQs did not improve the veracity detection accuracy of interviewers. The use of UQs did improve the accuracy of independent observers. However, this effect was mainly due to increased accuracy rates when judging the transcripts of interviews that had used Spatial and Temporal UQs as opposed to the Planning-based UQs. Moreover, the two different forms of UQs were shown to elicit qualitatively distinct forms of information in the interviewees' responses. Finally, the findings failed to support the cognitive load theory, with no interaction found between veracity and question type. Taken together, the findings of Chapters 2 and 3 suggest that, while unexpectedness may provide a useful dimension by which to assess the quality of interview questions, unexpectedness alone is not sufficient in capturing question quality.

Chapter 4 comprised an exploratory investigation into question generation. Novice participants were required to generate investigatively useful questions in a study which manipulated the topic of interview, the temporality of the topic and access to training. Questions were rated for general quality and as well for novelty and utility, the two dimensions by which creativity is generally judged (Finke, 1990). Findings revealed that the creativity dimensions were not a reliable method of question rating; experts showed little agreement in their ratings. Using the general quality ratings instead, the results indicated that there was an interaction effect between the topic of interview and temporal perspective on the quality of the questions generated. We argued that this finding was due to the scope of episodic knowledge inherently available to the interviewer, based on the topic/temporality combination.

In Chapter 5 a bottom-up approach was taken in order to establish further dimensions of question quality. A pilot study, in which a card sort was conducted on a

large set of generated questions, revealed three potential dimensions of question quality : Relevance, Unpredictability, and Knowledge probed. Following this, a second round of question generation was conducted, with questions being rated using this 3-dimensional model. The findings revealed that, when the three components of the model were summed, it was a reliable method for rating question quality. Additionally, scope was shown to be an important factor in one's ability to generate good-quality questions, though the effect may be context-dependent.

Finally, in Chapter 6 the 3-dimensional model was applied to real-world interview questions. Transcripts of aviation security interviews were rated using the model, as well as for general quality. The findings showed that ratings made using the 3-dimensional model were generally reliable and positively correlated with the general quality ratings. Moreover, they were effective in predicting the success of the interview outcome, with questions from successful transcripts being rated higher on each of the three dimensions than questions from the unsuccessful transcripts. We concluded that the 3-dimension model developed in Chapter 5 is a useful and reliable model by which to rate the quality of interview questions.

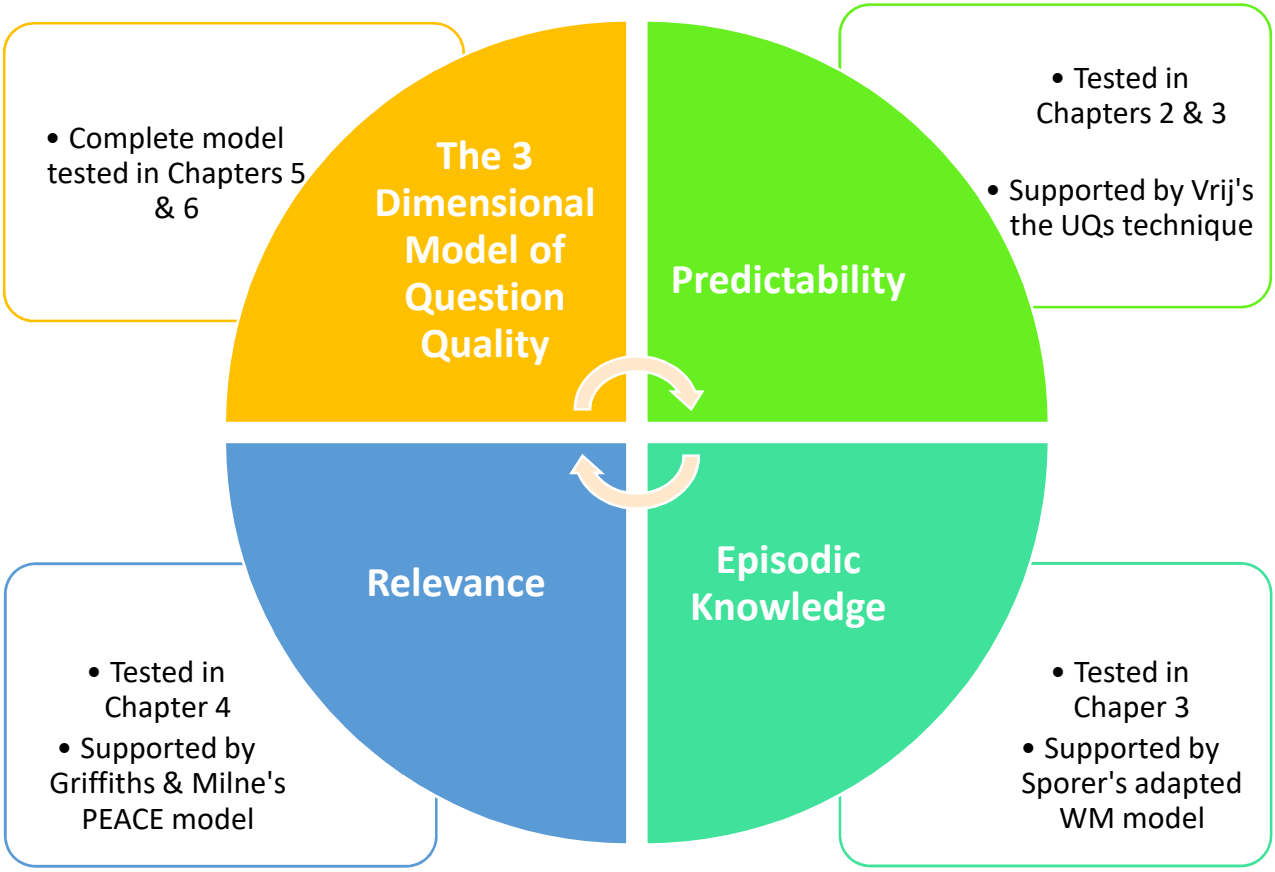

Figure 1. A visual representation of the studies and theories contributing to the development of the 3-dimensional model of question quality.

What makes a good quality interview question?

1) *What are the dimensions by which good- and poor-quality questions differ?*

2) *Is it possible to rate those dimensions objectively?*

Anticipation. The first step in addressing these research questions was to investigate the UQs technique (Vrij et al., 2009). There is a wealth of research which shows that when a question is not anticipated, this can reveal distinct patterns of verbal behaviours from honest and deceptive interviewees (Lancaster et al., 2013; Sooniste et al., 2013; Vrij, Mann, et al., 2012). As such, unexpectedness was considered a natural starting point in the search for reliable dimensions of question quality. The approach is

based on the assumption that liars will often attempt to anticipate what they might be asked in an interview and prepare plausible sounding responses (Vrij et al., 2017). This is referred to as a 'lie script' (Colwell et al., 2007). However, this tactic will only prove successful if they correctly anticipate the questions they are asked. As such, asking questions which they have not anticipated removes the option for them to use this lie script and forces them to 'think on their feet' and lie spontaneously. According to Vrij (2014), this increases the cognitive load of a liar, but not a truth-teller as they are able to rely on their memory of the event in question. Across a number of studies, it has been shown that the UQs approach to investigative interviewing is effective in eliciting verbal cues to deception (Vrij, Leal, et al., 2012; Vrij et al., 2011).

In Chapter 2, a systematic review of the studies investigating the UQs approach was undertaken. In total, 13 papers were identified that the met the inclusion criteria, comprising 16 experiments. The findings of the review showed that, on average, liars' responses to UQs were shorter, less detailed and less plausible than truth-tellers' responses. This provided support for the approach and suggests that it is an effective method for eliciting cues to deception. As such, this indicates that unexpectedness could be a vital dimension by which to judge the quality of interview questions; the more unexpected a question is, the more likely it is to reveal some investigatively useful information.

However, the review uncovered some issues with the 13 papers. Firstly, no study had investigated the effect of UQs on interviewer veracity detection accuracy. As such, it was not clear from the current research whether using UQs would have a beneficial effect on an interviewer's ability to detect deceit. Two of the experiments had measured observer accuracy, whereby participants are shown transcripts of the interviews and asked to judge veracity (Vrij, Leal, et al., 2011, 2012). However, there are situations,

such as aviation security screening, where it is vitally important that the interviewer is able to determine the veracity of the interviewee. Secondly, there was an abundance of different types of UQs used across the studies. These included UQs that focussed on planning, spatial details, temporal details, and sketches (Clemens et al., 2013; Lancaster et al., 2013; Vrij et al., 2011). Critically, no study included in the analysis had directly compared these various question types. Finally, there was only mixed support for the cognitive load theory. Some studies had shown that liars found the UQs more cognitively demanding than truth-tellers (e.g., Mac Giolla & Granhag, 2015). However, others had shown that they found them equally difficult (Sooniste et al., 2013, 2014, 2015).

In Chapter 3, a large-scale two-part experiment was devised in order to further investigate these issues. Participants were asked to either complete a navigation task (truth-tellers) or pretend they had completed the task (liars) and were subsequently interviewed. The interviews comprised a set list of questions that were either general questions about the task that we expected the interviewees to anticipate being asked (Anticipated), questions about the planning of the task (Unanticipated/Planning), or questions about the specific spatial and temporal details associated with carrying out the task (Unanticipated/Spatial and Temporal). Following each interview, the interviewee was required to state how difficult they had found answering the questions and the interviewers were asked to state whether they thought the interviewee was telling the truth or not.

The results showed that the use of UQs did not improve interviewer veracity detection accuracy. Additionally, there was no interaction between veracity and question type on the reported difficulty experienced in answering the questions. Transcripts of the interviews were shown to a separate group of participants who were

asked to determine the veracity of the interviewee. These results showed that UQs did improve observer accuracy, though this effect was much greater for the transcripts of the Spatial and Temporal UQs than the Planning UQs. Moreover, a Reality Monitoring analysis revealed that the two forms of UQ resulted in qualitatively distinct information being gathered in response. Planning UQs led to more cognitive operations words, whilst the Spatial and Temporal UQs led to more episodic information words. In theory, responses that are rich in episodic details should be more beneficial in terms of distinguishing between honest and dishonest accounts than responses that are rich in cognitive operations words (Oberlader et al., 2016; Sporer, 2016; Vrij et al., 2007).

Taken together, the findings of Chapters 2 and 3 suggest that the unexpectedness of a question is a potentially valuable dimension by which to judge question quality. Asking questions that an interviewee does not anticipate being asked can lead to differences in the verbal behaviours of truth-tellers and liars. These differences appear to be noticeable to independent observers who view transcripts of the interviews. This is important for investigative questions, given that it is independent observers, such as jury members or judges, who are often tasked with determining the veracity of an individual's account, based on their interview statement (Haworth, 2018). However, such differences do not seem to be noticeable to the interviewer, during the interview itself. Furthermore, given the findings that different types of UQ can have different effects on interview outcomes and gather distinct forms of information, it is clear that anticipation alone is not sufficiently able to capture question quality; the type of knowledge sought by the questions may be equally important.

Creativity. The next approach taken in determining the dimensions of question quality was to investigate creativity. Looking at the CCE interview technique (Ormerod & Dando, 2015), the questions promoted are unexpected tests of expected knowledge.

Generating these questions is an inherently creative skill. Therefore, building on the work in Chapters 2 and 3, we explored creativity in Chapter 4, investigating whether it might be a useful approach in identifying the dimensions of question quality. The standard definition of creativity involves two distinct dimensions: novelty and utility (Finke, 1990). Whilst novelty is most commonly associated with creativity, a unique product is not considered truly creative unless it is also effective in achieving its intended purpose (Runco & Jaeger, 2012). This definition is arguably applicable to the quality of interview questions; a good quality question will be one that is both novel (i.e., not likely to be anticipated by others) and useful (i.e., generates some investigatively relevant information). Therefore, an experiment was devised to investigate the use of these two dimensions and determine their applicability to question quality.

In chapter 4, novice participants were shown a series of information gathering interview clips in which four interviews discussed four broad topics (home, hobbies, travel and work) from either a past or a present temporal perspective. Following each video, the participants were required to generate an investigatively useful follow-up question. Five experts rated the questions for both novelty and utility. Additional, two of the experts rated the questions using a more subjective, implicitly-defined measure of general quality. The results showed that there was no agreement between the expert judges in terms of novelty or utility. Despite being objectively and clearly defined, the experts appeared to hold differing views regarding the novelty and utility of interview questions, and the effect was not mediated by their professional background (academic or practitioner). Moreover, looking at the individual correlations between the two judges' ratings who had applied both rating techniques showed that, for both, there was a correlation between utility and general quality, but not between novelty and general

quality. This suggests that generating useful information is more important to question quality than the originality of the question.

Taken together, it was concluded that the two dimensions of creativity were not reliable scales by which to judge question quality. At this point the best model for judging question quality was a simple, subjective measure of global quality. Research in the creativity field has promoted the use of such implicitly-defined global ratings, arguing that it allows the raters to apply their own criteria to the judgement and often results in more reliable findings (Amabile, 1982). However, this is not prescriptive enough for the purposes of rating test questions. It does not provide useful information to the investigative community regarding the specific dimensions that underly the quality of interview questions and does not offer an objective model with which to further explore the factors that potentially influence one's ability to generate good-quality questions. Therefore, in Chapter 5 we sought to address this by using a bottom-up approach in order to develop an objective model by which to judge question quality.

Applying a bottom-up approach.

Card Sort. In Chapter 5 we employed a card sort technique in order to establish potential dimensions that are capable of separating good-quality and poor-quality questions. In an initial pilot study, novice participants were shown information gathering interview clips concerning a specific event. After each clip they were asked to generate an investigatively useful question. A bottom-up approach was taken to sort the questions into groups that had similar properties. This technique established seven distinct question categories, each of which instinctively varied in terms of their investigative value. Subsequent evaluation of these categories led to the identification of three dimensions by which each of the categories varied: relevance, unpredictability and

the type of knowledge probed by the question. By assigning each question category with a low, medium or high ranking for each of the three dimensions, it was demonstrated that each category could be distinguished without overlap. As such we theorised that the three dimensions would be effective in determining question quality and this formed the basis for a new rating model.

The 3-dimensional model of question quality. The three dimensions identified by the card sort were relevance, unpredictability and type of knowledge probed. The investigative relevance of questions is arguably intuitive. For example, an individual could generate a question that is tapping into episodic knowledge and is completely unpredictable, but if the information it is seeking is of no relevance to the investigation in hand then it has limited value. For example, the PEACE model of interviewing encourages police officers to plan and prepare their interviews, determining the relevant points needed for consideration. Unpredictability builds on the work of Vrij and colleagues (2009), as well as the findings of Chapters 2 and 3. Whilst we argued that unexpectedness alone is not sufficient in capturing question quality, its value in terms of eliciting verbal cues to deception has been demonstrated repeatedly (Lancaster et al., 2013; Vrij et al., 2009; Vrij, Mann, et al., 2012). Finally, the knowledge dimension refers to the type of knowledge probed by the question; is the question seeking general semantic knowledge that does not require specific experience of the event in question, or is it probing specific episodic knowledge that does require a genuine experience of the event? We argue that questions which probe episodic knowledge are more useful than those which probe semantic knowledge. Recalling episodic knowledge is more cognitively demanding (Taylor & Dando, 2018) and should lead to differences between truth-tellers' and liars' verbal behaviour (Dando et al., 2015).

In the main experiment in Chapter 5, we examined the reliability of the model in judging question quality. Participants were shown the same interview clips used in the pilot study and were asked to generate an investigatively useful question after each. Each question was subsequently rated by a group of experts using the 3-dimensional model, as well as for general quality. The results were mixed. There was an acceptable level of agreement found for the ratings of the 3-dimensional model as a whole. Also, there was an acceptable level of agreement between the experts' ratings of the knowledge dimension. However, there was only moderate agreement between the ratings of the relevance and unpredictability dimensions, neither of which reached the pre-determined threshold.

The results were also mixed in terms of the correlation between ratings on the 3-dimensional model and the general quality ratings. Three of the four judges' ratings on the model overall had a medium-to-strong correlation with their own general quality ratings, though one judge's ratings were not correlated at all. In terms of the individual dimensions, knowledge correlated most strongly with general quality for two of the judges and relevance correlated most strongly for the other two. However, unpredictability did not correlate with general quality ratings for any of the judges.

Taken together, the findings of Chapter 5 provided an exploratory first step in the development of a new dimensional model for judging the quality of investigative interview questions. Initial support for the model was mixed and concerns were raised with regard to reliability. Therefore, in order to further investigate the use of the model, we next applied it to a set of real-world interview questions, with genuine outcomes. Additionally, the mixed findings regarding reliability suggested that it might prove beneficial to provide raters with a more detailed, thorough explanation of the three dimensions in subsequent studies.

Testing the 3-dimensional model on real-world questions. In Chapter 6, transcripts were taken from the CCE interviews originally conducted in Ormerod and Dando's (2015) aviation security field-study. Half of the transcripts were of security screening interviews where a mock passenger had been successfully identified, and half were transcripts of interviews where a mock passenger had not been identified. This provided us with a large set of real-world questions, from a professional context, that had a genuine outcome measure by which to judge the reliability and efficacy of the 3-dimensional model. The questions were rated by two experts who had attended a brief training exercise in how to apply the model, in order to establish whether increased agreement would be found when the individual dimensions were more thoroughly explained.

The results suggested that this was the case; there was a strong agreement between ratings on the model overall, using the average rating of the three dimensions. Additionally, there was strong agreement between the ratings for the unpredictability and knowledge dimensions. However, there was low agreement found for the ratings of the relevance dimension. This is likely due to the context of the interviews. The interviews used in this study were aviation security screening interviews and, in this context, almost any question which challenges the presented identity of the interviewee will be relevant. Therefore, relevance was, arguably, less discriminatory than in a forensic interview regarding a specific incident. Overall, the findings of Chapter 6 represent an improvement in the reliability of the 3-dimensional model, suggesting the brief training exercise was effective.

Looking at the correlation between the 3-dimensional model and general quality, there was a strong positive correlation found between the average ratings across the three dimensions and the general quality ratings. Moreover, there was a positive

correlation found between the ratings on each of the three individual dimensions and general quality. This suggests that the 3-dimensional model was effectively capturing differences in the subjectively perceived quality of the questions. This was supported by the findings that showed that the questions taken from the successful interviews were rated higher on the average score on the model, as well as for each of the three individual dimensions, than questions taken from the unsuccessful interviews. This suggests that each dimension, in isolation, is able to capture some element of question quality that is sufficient in predicting the outcome of an interview.

Summary. Overall, this thesis has established that unexpectedness is a potentially important dimension of question quality and that UQs can help to distinguish between truth-tellers and liars (Lancaster et al., 2013; Vrij et al., 2009). But anticipation alone is not sufficient in accounting for these effects; the type of UQ can affect both the type of information received in response and the outcome of the interview. The technique is more effective when the questions are both unanticipated and probing episodic knowledge. This is similar to the CCE approach, which has been shown to be an extremely successful technique for detecting deception (Ormerod & Dando, 2015). Whilst the CCE technique requires a reasonable level of creativity on the part of the interviewer, the standard components of creativity (novelty and utility; Finke, 1990), did not prove to be reliable dimensions by which to rate question quality. The subjective, implicitly-defined measure of global quality was reliable, but did not answer the questions we set out to investigate and has limited use in investigative interviewing research.

In order to address this, we used a bottom-up approach to identify a 3-dimensional model which we theorised could be reliably used to distinguish between good- and poor-quality questions. The three dimensions were relevance,

unpredictability and the type of knowledge probed. Across two studies, the reliability of the new model was tested, receiving mixed support. Inter-rater agreement was questionable in Chapter 5, but an improvement was noted when direct, thorough explanation of the three dimensions was provided in Chapter 6. This indicates that further work is required in terms of defining the three dimensions, and how those definitions are conveyed, given that the training conducted in Chapter 6 was time consuming and potentially unfeasible for future studies. However, the model did appear to encapsulate question quality. There was a positive correlation between ratings on the three dimensions and ratings of subjective general quality. Moreover, the three dimensions were effective in predicting the success of interview outcomes.

Of course, establishing a method by which to assess the quality of an interview question does not, in itself, provide a good investigative interviewer. There are many important factors to consider if we were instead attempting to judge the overall quality of the interviewer, such as rapport building and empathy, both of which can play important roles in the potential success of the investigative interview (Bull & Cherryman, 1995; Griffiths & Milne, 2006; Walsh & Bull, 2010). However, the findings taken together from this thesis do suggest that the 3-dimensional model is a potentially useful tool by which to judge the quality of interview questions and does allow for subsequent exploration into the factors affecting question generating ability. As such, we feel it makes a positive contribution to the investigative field.

Factors Affecting Question Quality

3) *What factors enhance or inhibit one's ability to generate good-quality questions?*

Scope. In Chapter 4 the results showed that there was an interaction effect between the topic of interview and the temporal perspective of the topic on the quality of subsequently generated questions. Follow-up tests showed that when the topic of the interview was Travel, higher quality questions were generated when the topic was framed in a past temporal perspective than a present perspective. Conversely, when the topic was Work, higher quality questions were generated when the topic was framed in a present temporal perspective. We argued that this was due to the scope of available episodic information inherently associated with the topic/temporality combination. Taking the topic of travel, for example, asking questions about previously experienced travel excursions allows the interviewer to explore a rich vein of episodic memory, if the interviewee has genuinely experienced that excursion (i.e., there is a broad scope). In contrast, for a present excursion which is due to take place, the interviewer is limited to exploring the planning of that excursion and as such there is less episodic information available (i.e., there is a narrow scope).

As has been stated previously in this discussion, honest accounts of an event tend to be rich in information associated with episodic memory, such as the visual, spatial or temporal details encountered during an incident (Vrij et al., 2007). Whereas, imagined accounts tend to include more references to internal thought processes, due to the fact that the experience of the event has been conceived endogenously (Oberlader et al., 2016). As such, it is reasonable to assume that discussions in which there is a broad scope for episodic inquiry will provide more opportunities to ask investigatively useful questions and, in turn, lead to more positive interview outcomes. In Chapter 5 we explored this theory further by manipulating the scope of episodic information. Participants conducted a series of tasks, half with specific, rigid instructions (narrow scope) and half with more general, flexible instructions (broad scope). They were

subsequently interviewed about their involvement in the tasks. Videos of those interviews were shown to a group of novice and expert participants who were asked to generate an investigatively useful question after each. Finally, a group of experts rated the questions using the 3-dimensional model.

Results indicated that the scope of episodic information available did affect the quality of the subsequently generated questions. However, this finding went against our original prediction; the tasks which had a narrow scope led to the generation of higher rated questions on the 3-dimensional model than the tasks which had a broad scope. This refuted our prediction that the broader the scope of episodic information, the greater the opportunity to generate high-quality questions. Moreover, looking at the ratings given on the three individual dimensions revealed that this effect was only found for the knowledge ratings. Therefore, the narrow scope tasks led to the generation of questions which were rated as probing more episodic knowledge than the broad scope tasks. Whilst this contrary finding refuted our original theory, it can perhaps be explained in terms of the context of the interview.

In Chapter 4, when the theory was originally developed, the interviews comprised informal discussions concerning a series of broad topics, similar to the interviews conducted in Ormerod and Dando's (2015) aviation security screening study. In contrast, in Chapter 5 the interviews concerned a series of specific incidents and were designed to more closely resemble a forensic context. The scope of episodic information available was shown to have an effect on subsequent question generation in both experiments, but in opposite directions. As such, we can conclude that scope is an important factor which can affect one’s ability to generate good-quality interview questions, though the context of the interview is potentially crucial. A broad scope is useful when the context is involves broad discussions of general topics; however, in

more forensic contexts where a specific incident is under investigation, a narrow scope might provide a more rigid framework with which to examine specific episodic details.

Training. In Chapters 4 and 5 the majority of the question generating participants were novices. This is a limitation of the present thesis. We were attempting to understand the conditions and factors affecting the generation of investigative interview questions and in professional contexts the individuals who are generating these questions have usually been trained to a high level. For example, the PEACE model of interviewing requires a week-long training course; the advance model requires a three-week course (Griffiths & Milne, 2006). The security agents in Ormerod and Dando's (2015) field study received one week of classroom training, followed by one week of on-the-job training, when learning CCE. As such, the novice participants in Chapters 4 and 5 were unlikely to contribute questions that were as high in quality as trained professionals. Despite this, novices were still recruited, due to practicality. One advantage to this was that we could assess the effect of a brief training intervention on individuals' ability to generate interview questions.

In order to investigate the effect of training, we developed a video designed to improve the question generating ability of novice participants. The video was 10-minutes long and covered the basic principles of the CCE technique, with a focus on the final veracity-testing phase. It included sample clips that showed an individual both telling the truth and lying about a holiday in order to demonstrate how the interviewer conducts each step of the process. In Chapter 4, the use of this video was manipulated so that half of the novice question generators viewed it before taking part in the task and half had the option of watching it after taking part in the task. The results showed that the participants who had watched the video before taking part generated higher quality questions than those who had not.

This finding gave us confidence that the training video was an effective tool for improving the performance of novice question generators. It also provided support for the efficacy of the CCE technique given that it was able to improve performance even from such a brief introduction to the method. Additionally, it suggests that the principles of CCE are simple to convey and easy to understand. Nonetheless, it should be noted that the trained participants in Chapter 4 had an average quality rating around the mid-point of the scale. Whilst the training improved their performance, it is clear that more advanced trained is required to generate questions that are above average in quality. Overall, the findings suggest that training is a factor that affects one's ability to generate good-quality interview questions. Arguably, given the creative nature of question generation, one might reasonably expect to find that some are more naturally gifted in this skill than others. Nevertheless, the results of Chapter 4 indicate that training may supersede individual differences in ability and has a positive effect on ability.

Veracity. The veracity of an interviewee has been shown to affect verbal behaviour. DePaulo and colleagues' (2003) meta-analysis revealed that statements given by deceptive interviewees tend to be less consistent, less coherent and less detailed than those given by truth-tellers. This finding has been consistently supported in empirical studies (Bogaard et al., 2019; Hartwig et al., 2007; Vrij et al., 2018). In Chapter 3 we showed that questions which gather accounts rich in episodic information lead to higher veracity detection accuracy than those which gather accounts rich in cognitive mechanism information. This is likely due to the fact that liars do not have access to the specific episodic information, given that they have not actually experienced the event in question, and therefore their accounts tend to be low in episodic detail (Sporer, 2016). But there is no research examining whether the veracity of an interviewee affects the questions generated by interviewers. It's feasible that these differences in verbal

behaviour, exhibited by truth-tellers and liars, affect the information available to interviewer which in turn affects their question generating ability.

In Chapter 5, interviewees were asked to conduct four separate tasks. Veracity was manipulated so that in half the tasks they were asked to conduct them honestly, and in half they were asked to conduct them deceptively. Furthermore, a distinction was made between two forms of deception: hiding and pretending. Memon and colleagues (2013) conducted a literature review which showed the difference that these two forms of lying can have on interview outcomes. They argue that hiding an act that one has committed is more cognitively demanding than pretending an act has taken place, given that hiding requires the deceiver to simultaneously hold two mental representations. The results of Chapter 5 showed that veracity had an effect on novices' ratings of general quality. Interestingly, there was no difference in quality found between questions generated for the honest and deceptive interviews in general, though there was a difference in quality between the two forms of lying; questions generated for interviews when the interviewee was hiding were rated higher in quality than those generated for interviews when the interviewee was pretending.

These findings support the unpublished review conducted by Memon and colleagues (2013). The dual mental representation required for hiding the truth has been shown to require a greater level of cognitive effort than single mental representations (Ormerod & Richardson, 2003). The UQ and CCE interview techniques both rely on differences in cognitive load faced by truth-tellers and liars (Ormerod & Dando, 2015; Vrij, 2014). However, the cognitive load theory might not be so straight-forward; given that different forms of lying apparently lead to distinct verbal behaviours. In turn, this distinction between the two forms of deception appears to affect one's ability to generate good-quality questions. This suggests that the increased cognitive load faced

by hiders affects their verbal behaviour in a manner that allows an interviewer to generate more investigatively useful questions. As such, future research should perhaps examine methods for preventing deceptive interviewees from simply pretending and try to constrain their account in such a way that requires them to hide the truth.

Expertise. The expertise of the question generators was also manipulated in Chapter 5. It is intuitive to assume that experts will perform better than novices, in any given domain. There is research to support this idea, showing that experts tend to demonstrate wider knowledge and have better problem-solving ability in their specific domain of expertise (Klein & Hoffman, 1992; Mosier et al., 2018). Fahsing and Ask's (2016) investigation into the ability of experienced and novice police officers revealed that the more experienced detectives generated more alternative hypotheses than the novice officers when presented with a hypothetical missing person case. In order to investigate the effects of expertise, Chapter 5 recruited a group of experts to generate questions, as well as a group of novices. The results revealed that expertise had no effect on the ratings made by other experts. However, there was an effect on the ratings made by novices; questions generated by experts were rated higher in quality than those generated by novices.

The finding that expertise affects the ratings of novices, but not experts, is supported by the findings of Runco and colleagues (1994). In their examination of the creativity ratings made by expert and novice judges, they concluded that experts tend to be overly critical and, in turn, less sensitive to differences in ability. Likewise, expert and novice raters in Chapter 5 appeared to value distinct facets of question quality. Future research may wish to examine a sample of the questions that novices rated high in quality, as well as a sample of questions that experts rated highly, and conduct a qualitative content analysis of the questions. This may reveal the distinct patterns of

language or content that separated their opinions and, subsequently, help to establish the individual dimensions of question quality.

It should be noted the term ‘expert’, as used in Chapter 5, is a potential source of debate. The definition applied in that chapter was “any individual with five or more years’ experience in an investigative role.” However, in reality, this definition does not guarantee expertise, but merely experience. The term was used in order to distinguish these participants from the novices and, in turn, to look at the effects of experience on question generating ability. Therefore, the conclusions drawn, with regards to expertise, should be treated with caution. Subsequent studies could attempt to drill down further into the effect of expertise, perhaps by using only participants who have achieved the PEACE advanced level interview training, in an effort to elicit a higher quality of generated questions. In turn this would allow for a more thorough investigation into the potential factors which lead to higher quality questions being generated.

Application of Current Findings

Arguably, the findings presented in this thesis are of most relevance to the CCE technique. The 3-dimensional model, developed in Chapter 5, was shown to be able to predict the outcome of CCE interviews in Chapter 6. As such, it has been demonstrated that, specifically when using the CCE technique, a question is more likely to result in a successful outcome if it is relevant, unpredictable, or probes episodic knowledge. It is most likely to be successful if it encompasses all three of those dimensions. As such, this should be taken into consideration by researchers who are continuing to develop the CCE technique, as well as the experts who are currently training the technique to those in the investigative community. Training needs to promote the use of veracity test questions which are relevant, unpredictable and probe episodic knowledge. In Chapter 4

we showed that the principles of the CCE technique can be conveyed succinctly, in a brief 10-minute video, leading to improvements in question generation. Future research may wish to explore the use of this training video further in order to establish how effective it is in training the CCE technique in a more professional context. If it can be demonstrated that CCE is able to be effectively learnt remotely, this would be a valuable attribute of the technique. Benson and Powell's (2015) study into the efficacy of a predominantly web-based training program, designed to improve forensic interviews with children, showed that remote learning can be effective.

Whilst the findings presented in this thesis are most relevant to the CCE technique, they are still relevant to the wider realm of investigative interviewing. The PEACE model of interviewing instructs officers to plan and prepare their interviews (Griffiths & Milne, 2006). This includes making a written plan that includes the points necessary to prove an offence and topics to be covered. Proper planning has been shown to improve the chances of gaining a positive interview outcome (Walsh & Bull, 2010). Based on the findings presented in this thesis, part of this planning should include consideration of how to phrase important questions in order to ensure that they are relevant, unpredictable and probing episodic information. This may be difficult to do in advance, as questions often arise from details in the interviewee's account. However, considering the SUE technique (Granhag et al., 2007), whereby critical evidence is initially withheld from an interviewee and revealed strategically during the course of the interview, it is feasible that prior consideration regarding the phrasing of important questions is possible, perhaps whilst strategically revealing pieces of evidence.

This thesis has explored both the value of unanticipated questions, and the extent to which the factors that underlie the creation of high-quality unanticipated questions can be identified. Although the research is a fundamental empirical exploration rather

than the development of an interviewing system per se, in future work the extent to which the three dimensions identified as underlying good question design can be trained might be investigated. It seems likely, for instance, that in information gathering interviews, training in good question design will impact positively on the amount and veracity of information gathered. The approach has less application in interrogation contexts typically found in the US judicial system, where a confession is the desired outcome. However, training based on the principles identified across these studies would be entirely consistent with the UK PEACE approach to interviewing, which many countries are adopting as best practice.

When discussing the potential of incorporating these findings into any training procedure, one thing to consider is the extent to which skills generally transfer from the classroom to the real world. There is evidence from the investigative interviewing domain to suggest that this is often not the case. The initial examination of PEACE training (Mcgurk et al., 1993) showed that, 6 months after training, the officers who had received it were still applying those skills which they had been taught. However, more recently, Clarke and colleagues (2011) found that the only difference in the interviewing carried out by PEACE-trained and non-trained officers was the length of the interview, with no significant differences found in terms of the core skills associated with PEACE training. In contrast, Griffiths and Milne (2006) showed that officers who had received the 3-week advance training had an increased level of interviewing skill when returning to the workplace. Although, the level of some of the more complex skills was shown to drop after around one year. As such, future research which investigates methods for incorporating the findings presented in this thesis into investigative interview training, should first consider focusing on how to ensure that such skills are capable of being transferred from the classroom to the interviewing room.

Additionally, the thesis highlights the need for a multi-dimensional approach to investigative interviewing. For example, the results of Chapters 2 and 3 suggest that asking UQs can elicit differences in truth-tellers' and liars' verbal behaviour. However, anticipation alone did not account for these differences. Chapter 3 showed that UQs which focuses on episodic information were more useful and led to increased observer veracity detection accuracy than UQs which focused on planning. Therefore, in order for UQs to be a successful technique for veracity detection, other considerations, such as the type of information probed, must be considered. This is perhaps the reason for the efficacy of the CCE technique in Ormerod and Dando's (2015) field study, given that CCE encompasses six empirically tested techniques.

One potential future application of the findings involves interesting new technological developments. For some time, there has been research exploring the use of technology in investigative interview contexts, though this tends to focus on collecting witness reports. For example, Chih and colleagues (2008) developed a system, using natural language processing (NLP), that is designed to allow an individual to report a crime anonymously online. They claim that, by using NLP, the system is capable of extracting crime-relevant information, which it subsequently uses to generate necessary follow-up questions. More recently, Shih, Chen, Syu, and Deng (2019) have proposed a cloud-based online crime reporting system that securely protects the informer's identity.

However, as yet, this type of technology has not been applied to suspect interviews or used to help interviewers. Early-stage research at the University of Sussex is currently exploring this concept. Work is being undertaken to develop an application which, using NLP, will be able to listen to an interview in real time, extract key details from the interviewee's account and gather information regarding those details, which

may be able to assist the interviewer. The long-term aim of this research project is for the application to be able to suggest questions to the interviewer. However, any such system will require a set of guidelines by which to rank the information that has been gathered, so that only the most useful information is shown to the interviewer. The 3-dimensional model developed in this thesis could help to inform the developers of this technology and be utilised as the basis for this ranking.

Future Directions

In Chapter 5 we investigated question generation from interviews that were forensic in nature and focused on a specific series of incidents. However, these incidents were staged and the illicit activity in questions was somewhat arbitrary in nature. Furthermore, there was no outcome measure in terms of the success of the interviews. In Chapter 6, we investigated genuine interview questions from a professional context, which did have an outcome measure by which to judge the success of the interview. However, the interviews themselves were not forensic in nature and involved general discussion of broad topics.

Further research is needed to combine these methodologies and investigate question generation, and the efficacy of the 3-dimensional model, using genuine questions from a forensic context with real outcome measures. Similar research has been conducted previously. For example, Leahy-harland and Bull (2017) examined transcripts of genuine suspect interviews in order to explore the strategies employed by police interviewers. Similarly, Walsh and Bull (2010) investigated the use of each component of the PEACE model in genuine suspect interviews. Both of these studies devised a measure by which to judge the success of an interview. A similar approach

could be applied in future research using the 3-dimensional model, with questions taken from genuine suspect interviews being rated in each of the dimensions.

In order to determine whether the 3-dimensional model is a useful tool that can be applied to interview training, a number of experiments could be devised. A training exercise could be developed that is focused on the 3-dimensional model. This could take the form of a short video, similar to the CCE training video employed in Chapters 4 and 5. Subsequently, a question generation study could be conducted in which half the participants had received this training and half had not. If the participants who had received training generated better quality questions than those who had not, this would provide further support for the theory that the three dimensions identified in this thesis encapsulate question quality. In a separate study, participants who had received this training could be asked to conduct interviews with truth-tellers and liars, as opposed to simply generating follow-up questions. This would examine the extent to which the 3-dimensional model can be trained, how well it is adhered to post-training and whether it improves veracity detection accuracy.

Conclusion

In summary, the findings presented in this thesis reveal that it is possible to establish the dimensions underlying the inherent quality of investigative interview questions. The 3-dimensional model, proposed in this thesis, has been shown to be a potentially reliable method for distinguishing between good- and poor-quality questions. Moreover, it was shown to be capable of predicting the outcome of an investigative interview. However, this is still a preliminary model and further examination will be required in order to determine its efficacy across a wider range of interview contexts. Additionally, the findings of this thesis have demonstrated that there

are numerous factors that might enhance or inhibit one's ability to generate an investigatively useful interview question, such as the veracity of the interviewee or the expertise of the question generator. One of the more intriguing of these factors is the scope of episodic knowledge inherently available to the interviewer, which was shown to have a context-dependent effect on question generating ability. The findings presented in this thesis have important implications for researchers who are continuing to develop interview techniques, such as the CCE approach, and also for investigative practitioners. Both groups are advised to pay careful consideration to the extent to which the questions they ask in interviews are relevant, unpredictable and probing episodic knowledge; as the findings of this thesis suggest that these three dimensions in combination should result in investigatively useful questions that lead to successful interview outcomes.

References

Aamondt, M. G., & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *The Forensic Examiner*, *15*, 6–11. Retrieved from http://www.questia.com/library/1G1-142682689/who-can-best-catch-a-liar-a-meta-analysis-of-individual

Abbe, A., & Brandon, S. E. (2013). The Role of Rapport in Investigative Interviewing: A Review. *Journal of Investigative Psychology and Offender Profiling*, *10*(3), 237–249. https://doi.org/10.1002/jip.1386

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, *43*(5), 997–1013. https://doi.org/10.1037/0022-3514.43.5.997

Authorised Professional Practice. (2019). Investigative Interviewing. Retrieved from https://www.app.college.police.uk/app-content/investigations/investigative-interviewing/

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, *63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baldwin, J. (1993). Police Interview Techniques: Establishing Truth or Proof? *The British Journal of Criminology*, *33*(3), 325–352. Retrieved from https://academic.oup.com/bjc/article/33/3/325/386571

Benson, M. S., & Powell, M. B. (2015). Evaluation of a comprehensive interactive training system for investigative interviewers of children. *Psychology, Public Policy, and Law*, *21*(3), 309–322. https://doi.org/10.1037/law0000052

Blair, J. P., & Kooi, B. (2004). The Gap between Training and Research in the Detection of Deception. *International Journal of Police Science & Management*, *6*(2), 77–83. https://doi.org/10.1350/ijps.6.2.77.34465

Blair, J. P., Levine, T. R., & Shaw, A. S. (2010). Content in context improves deception detection accuracy. *Human Communication Research*, *36*(3), 423–442. https://doi.org/10.1111/j.1468-2958.2010.01382.x

Bogaard, G., Colwell, K., & Crans, S. (2019). Using the Reality Interview improves the accuracy of the Criteria-Based Content Analysis and Reality Monitoring. *Applied Cognitive Psychology*, 1–14. https://doi.org/10.1002/acp.3537

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, *19*(3), 313–329. https://doi.org/10.1002/acp.1087

Bull, R., & Cherryman, J. (1995). *Helping to indentify skills gaps in specialist investigative interviewing: Enhancement of professional skills literature review*. London: Home Office.

Chih, H. K., Iriberri, A., & Leroy, G. (2008). Crime information extraction from police and witness narrative reports. *2008 IEEE International Conference on Technologies for Homeland Security, HST'08*, (June), 193–198. https://doi.org/10.1109/THS.2008.4534448

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The contributions of prefrontal cortex and executive control to deception: Evidence from activation likelihood estimate meta-analyses. *Cerebral Cortex*, *19*(7), 1557–1566. https://doi.org/10.1093/cercor/bhn189

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cicchetti, D. V., & Sparrow, S. S. (1990). Assessment of adaptive behavior development in young children. In J. H. Johnson & J. Goldman (Eds.), *Developmental Assessment in Clinical Child Psychology* (pp. 173–195). New York: Pergamon Press.

Clarke, C., & Milne, R. (2001). *National Evaluation of the PEACE Investigative Interviewing Course*. Police Research Award Scheme (PRAS/149).

Clarke, C., Milne, R., & Bull, R. (2011). Interviewing suspects of crime: The impact of

PEACE training, supervision and the presence of a legal advisor. *Journal of Investigative Psychology and Offender Profiling*, *8*(2), 149–162. https://doi.org/10.1002/jip.144

Cleary, H. M. D., & Warner, T. C. (2016). Police Training in Interviewing and Interrogation Methods: A Comparison of Techniques Used with Adult and Juvenile Suspects. *Law and Human Behavior*, *40*(3), 270–284. https://doi.org/10.1037/lhb0000175

Clemens, F., Granhag, P. A., & Strömwall, L. A. (2013). Counter-Interrogation Strategies when Anticipating Questions on Intentions. *Journal of Investigative Psychology and Offender Profiling*, *10*(1), 125–138. https://doi.org/10.1002/jip.1387

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Science* (Second Edi). New York: Routledge Academic.

Collins, R., Lincoln, R., & Frank, M. G. (2002). The effect of rapport in forensic interviewing. *Psychiatry, Psychology and Law*, *9*(1), 69–78. https://doi.org/10.1375/pplt.2002.9.1.69

Colwell, K., Hiscock-Anisman, C., Memon, A., Rachel, A., & Colwell, L. H. (2007). Vividness and spontaneity of statement detail characteristics as predictors of witness credibility. *American Journal of Forensic Psychology*, *25*(1), 5–30.

Colwell, K., Hiscock, C. K., & Memon, A. (2002). Interviewing techniques and the assessment of statement credibility. *Applied Cognitive Psychology*, *16*(3), 287–300. https://doi.org/10.1002/acp.788

Colwell, L. H., Miller, H. A., Lyons, P. M., & Miller, R. S. (2006). The Training of Law Enforcement Officers in Detecting Deception: A Survey of Current Practices and Suggestions for Improving Accuracy. *Police Quarterly*, *9*(3), 275–290. https://doi.org/10.1177/1098611104273293

CREST. (2016). Interview Tactics: The Unexpected Questions Technique. Retrieved from https://crestresearch.ac.uk/resources/unexpected-question-guide/

Dando, C. J., Bull, R., Ormerod, T. C., & Sandham, A. L. (2015). Helping to sort the liars from the truth-tellers: The gradual revelation of information during

investigative interviews. *Legal and Criminological Psychology*, *20*(1), 114–128. https://doi.org/10.1111/lcrp.12016

Dando, C. J., Geiselman, R. E., MacLeod, N., & Griffiths, A. (2016). Interviewing Adult Witnesses and Victims. In G. Oxburgh, T. Myklebust, & R. Milne (Eds.), *Communication in Investigative and Legal Contexts: Integrated Approaches from Forensic Psychology, Linguistics and Law Enforcement* (pp. 79–107). Chichester: Wiley-Blackwell.

Debey, E., Ridderinkhof, R. K., De Houwer, J., De Schryver, M., & Verschuere, B. (2015). Suppressing the truth as a mechanism of deception: Delta plots reveal the role of response inhibition in lying. *Consciousness and Cognition*, *37*, 148–159. https://doi.org/10.1016/j.concog.2015.09.005

DePaulo, B. M., Malone, B. E., Lindsay, J. J., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

Dorfman, M. H. (1996). Evaluating the interpretive community: Evidence from expert and novice readers. *Poetics*, *23*(6), 453–470. https://doi.org/10.1016/0304-422X(96)00004-6

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177. https://doi.org/10.1037/1082-989X.1.2.170

Fahsing, I., & Ask, K. (2016). The making of an expert detective: the role of experience in English and Norwegian police officers' investigative decision-making. *Psychology, Crime and Law*, *22*(3), 203–223. https://doi.org/10.1080/1068316X.2015.1077249

Finke, R. A. (1990). *Creative Imagery: Discoveries and Inventions in Visualizations*. NJ: Lawrence Eilbaum Associates, Inc.

Frank, M. G., Yarbrough, J. D., & Ekman, P. (2006). Investigative Interviewing and the Detection of Deception. In T. Williamson (Ed.), *Investigative Interviewing: Rights, Research, and Regulation* (pp. 229–255). Devon, UK: Willan Publishing.

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The Illusion of Transparency:

Biased Assessments of Others' Ability to Read One's Emotional States. *Journal of Personality and Social Psychology*, *75*(2), 332–346. https://doi.org/10.1037/0022-3514.75.2.332

Glenberg, A. M., Schroeder, J. L., & Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition*, *26*(4), 651–658. https://doi.org/10.3758/BF03211385

Global Deception Research Team. (2006). A world of lies. *Journal of Cross-Cultural Psychology*, *37*(1), 60–74. https://doi.org/10.1177/0022022105282295

Graesser, A. C. (1981). *Prose Comprehension Beyond the Word*. New York: Springer.

Granhag, P. A., Mac Giolla, E., Sooniste, T., Strömwall, L., & Liu-Jonsson, M. (2016). Discriminating Between Statements of True and False Intent: The Impact of Repeated Interviews and Strategic Questioning. *Journal of Applied Security Research*, *11*(1), 1–17. https://doi.org/10.1080/19361610.2016.1104230

Granhag, P. A., Strömwall, L., & Hartwig, M. (2007). The SUE technique: The way to detect deception. *Forensic Update*, *88*, 25–29.

Granhag, P. A., & Vrij, A. (2005). Detecting Deception. In N. Brewer & K. D. Williams (Eds.), *Psychology and Law: An Empirical Perspective* (pp. 43–92). New York: The Guilford Press.

Griffiths, A., & Milne, B. (2006). Will it all end in tiers? Police interviews with suspects in Britain. In Tom Williamson (Ed.), *Investigative Interviewing: Rights, Research and Regulation* (pp. 167–189). https://doi.org/10.4324/9781843926337

Gudjonsson, G. H. (2003). Psychology brings justice: the science of forensic psychology. *Criminal Behaviour and Mental Health*, *13*(3), 159–167. https://doi.org/https://doi.org/10.1002/cbm.539

Gudjonsson, G. H. (2006). The psychology of interrogations and confessions. In T Williamson (Ed.), *Investigative Interviewing: Rights, Research and Regulation* (pp. 123–146). Devon, UK: Willan Publishing.

Gudjonsson, G. H., & Pearse, J. (2011). Suspect interviews and false confessions. *Current Directions in Psychological Science*, *20*(1), 33–37. https://doi.org/10.1177/0963721410396824

Hartwig, M., & Granhag, P. A. (2015). Exploring the nature and origin of beliefs about deception: Implicit and explicit knowledge among lay people and presumed experts. In P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Detecting deception: Current challenges and cognitive approaches* (pp. 125–153). Chichester: Wiley-Blackwell.

Hartwig, M., Granhag, P. A., & Luke, T. (2014). Strategic Use of Evidence During Investigative Interviews: The State of the Science. In D. C. Raskin, C. R. Honts, & J. C. Kircher (Eds.), *Credibility Assessment: Scientific Research and Applications* (pp. 1–36). https://doi.org/10.1016/B978-0-12-394433-7.00001-4

Hartwig, M., Granhag, P. A., & Strömwall, L. A. (2007). Guilty and innocent suspects' strategies during police interrogations. *Psychology, Crime and Law*, *13*(2), 213–227. https://doi.org/10.1080/10683160600750264

Hartwig, M., Granhag, P. A., Stromwall, L., Wolf, A. G., Vrij, A., & af Hjelmsäter, E. R. (2011). Detecting deception in suspects: Verbal cues as a function of interview strategy. *Psychology, Crime and Law*, *17*(7), 643–656. https://doi.org/10.1080/10683160903446982

Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *International Journal of Evidence and Proof*, *22*(4), 428–450. https://doi.org/10.1177/1365712718798656

Hedges, L. V, & Olkin, I. (1985). *Statistical Methods in Meta-Analysis*. London: Academic Press Inc.

Hekkert, P., & Van Wieringen, P. C. W. (1996). Beauty in the Eye of Expert and Nonexpert Beholders : A Study in the Appraisal of Art. *The American Journal of Psychology*, *109*(3), 389–407. Retrieved from https://www.jstor.org/stable/1423013

Hickey, M. (2001). Application of Amabile's Consensual Assessment Technique for Rating the Creativity of Children's Musical Compositions. *Journal of Research in Music Education*, *49*(3), 234–244. https://doi.org/10.2307/3345709

Inbau, F. E., Reid, J. E., Buckley, J. P., & Jayne, B. C. (2011). *Criminal interrogations and confessions*. Burlington, MA: Jones & Bartlett Learning.

Johnson, M. K., & Raye, C. L. (1981). Reality Monitoring. *Psychological Review*, *88*(1), 67–85.

Jundi, S., Vrij, A., Mann, S., Hope, L., Hillman, J., Warmelink, L., & Gahr, E. (2013). Who should I look at? Eye contact during collective interviewing as a cue to deceit. *Psychology, Crime and Law*, *19*(8), 661–671. https://doi.org/10.1080/1068316X.2013.793332

Kassin, S. M. (2006). A critical appraisal of modern police interrogations. In T Williamson (Ed.), *Investigative Interviewing: Rights, Research, and Regulation* (pp. 207–228). Devon, UK: Willan Publishing.

Kassin, S. M., Goldstein, C. C., & Savitsky, K. (2003). Behavioural confirmation in the interrogation room: On the dangers of presuming guilt. *Law and Human Behavior*, *27*(2), 187–203.

Kaylor-Hughes, C. J., Lankappa, S. T., Fung, R., Hope-Urwin, A. E., Wilkinson, I. D., & Spence, S. A. (2011). The functional anatomical distinction between truth telling and deception is preserved among people with schizophrenia. *Criminal Behaviour and Mental Health*, *21*, 8–20. https://doi.org/10.1002/cbm.785

King, W. R., & Dunn, T. M. (2010). Detecting deception in field settings: A review and critique of the criminal justice and psychological literatures. *Policing: An International Journal of Police Strategies & Management*, *33*(2), 305–320. https://doi.org/10.1108/13639511011044902

Klein, G. A., & Hoffman, R. R. (1992). Seeing the invisible: Perceptual-cognitive aspects of expertise. In M. Rabinowitz (Ed.), *Cognitive science foundations of instruction* (pp. 203–226). Mahwah, NJ: Eribaum.

Knieps, M., Granhag, P. A., & Vrij, A. (2013). Back to the Future: Asking About Mental Images to Discriminate Between True and False Intentions. *The Journal of Psychology*, *147*(6), 619–640. https://doi.org/10.1080/00223980.2012.728542

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lancaster, G. L. J., Vrij, A., Hope, L., & Waller, B. (2013). Sorting the Liars from the

Truth Tellers: The Benefits of Asking Unanticipated Questions on Lie Detection. *Applied Cognitive Psychology*, *27*, 107–114. https://doi.org/10.1002/acp.2879

Laville, S. (2012). Sam Hallam released after seven years in prison. *The Guardian*. Retrieved from https://www.theguardian.com/uk/2012/may/16/sam-hallam-released-seven-years

Leahy-harland, S., & Bull, R. (2017). Police strategies and suspect responses in real-life serious crime interviews. *Journal of Police and Criminal Psychology*, *32*(2), 138–151. https://doi.org/10.1007/s11896-016-9207-8

Levine, T. R., Blair, J. P., & Carpenter, C. J. (2018). A critical look at meta-analytic evidence for the cognitive approach to lie detection: A re-examination of Vrij , Fisher , and Blank (2017). *Legal and Criminological Psychology*, *23*, 7–19. https://doi.org/10.1111/lcrp.12115

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in Detecting Truths and Lies: Documenting the " Veracity Effect". *Communication Monographs*, *66*(2), 125–144. https://doi.org/10.1080/03637759909376468

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, *7*(4), 560–572. https://doi.org/10.1016/0010-0285(75)90023-7

Logue, M., Book, A. S., Frosina, P., Huizinga, T., & Amos, S. (2015). Using Reality Monitoring to Improve Deception Detection in the Context of the Cognitive Interview for Suspects. *Law and Human Behaviour*, *39*(4), 360–367. https://doi.org/10.1037/lhb0000127

Long, H. (2014). An Empirical Review of Research Methodologies and Methods in Creativity Studies (2003 – 2012). *Creativity Research Journal*, *26*(4), 427–438. https://doi.org/10.1080/10400419.2014.961781

Luke, T. J., Hartwig, M., Joseph, E., Brimbal, L., Chan, G., Dawson, E., … Granhag, P. A. (2016). Training in the Strategic Use of Evidence technique: Improving deception detection accuracy of American law enforcement officers. *Journal of Police and Criminal Psychology*, *31*(4), 270–278. https://doi.org/10.1007/s11896-015-9187-0

Mac Giolla, E., & Granhag, P. A. (2015). Detecting False Intent Amongst Small Cells

of Suspects: Single Versus Repeated Interviews. *Journal of Investigative Psychology and Offender Profiling*, *12*, 142–157. https://doi.org/10.1002/jip.1419

Maldarelli, G. A., Hartmann, E. M., Cummings, P. J., Horner, R. D., Obom, K. M., Shingles, R., & Pearlman, R. S. (2009). Virtual Lab Demonstrations Improve Students ' Mastery of Basic Biology Laboratory Techniques. *Journal of Mocrobiology & Biology Education*, *10*, 51–56. https://doi.org/10.1128/jmbe.v10.99

Mann, S., Ewens, S., Shaw, D., Vrij, A., Leal, S., & Hillman, J. (2013). Lying Eyes: Why Liars Seek Deliberate Eye Contact. *Psychiatry, Psychology and Law*, *20*(3), 452–461. https://doi.org/10.1080/13218719.2013.791218

Martonosi, S. E., & Barnett, A. (2006). How Effective Is Security Screening of Airline Passengers? *Interfaces*, *36*(6), 545–552. https://doi.org/10.1287/inte.l060.0231

Mcgurk, B. J., Carr, M. J., & Mcgurk, D. (1993). *Investigative Interviewing Courses For Police Officers: An Evaluation*. London: Home Office.

McKeithen, K. B., Reitman, J. S., Rueter, H. H., & Hirtle, S. C. (1981). Knowledge organization and skill differences in computer programmers. *Cognitive Psychology*, *13*(3), 307–325. https://doi.org/10.1016/0010-0285(81)90012-8

McKnight, L., Ormerod, T. C., Sas, C., & Dix, A. (2006). As useful as a bicycle for a fish: Exploration versus constraint in creativity. *Proceedings of the 26th Conference of the Cognitive Science Society, Vancouver*. Mahwah, NJ: LEA.

Memon, A., Fraser, J., Colwell, K., Odinot, G., & Mastroberardino, S. (2010). Distinguishing truthful from invented accounts using reality monitoring criteria. *Legal and Criminological Psychology*, *15*, 177–194. https://doi.org/10.1348/135532508X401382

Memon, A., Ormerod, T. C., & Dando, C. J. (2013). *How the task determines the effectiveness of reality monitoring in detecting deception*.

Memon, A., Vrij, A., & Bull, R. (2003). *Psychology and Law: Truthfulness , Accuracy and Credibility* (2nd Ed). Chichester: Wiley-Blackwell.

Milne, R., & Bull, R. (2003). Interviewing by the police. In D. Carson & R. Bull (Eds.), *Handbook of psychology in legal contexts* (pp. 111–125). Chichester: Wiley.

Milne, R., & Bull, R. (2016). Witness interviews and crime investigation. In D. Groome & M. W. Eysenck (Eds.), *An introduction to applied cognitive psychology* (2nd Ed, pp. 175–196). Oxford: Routledge.

Milne, R., Griffiths, A., Clarke, C., & Dando, C. J. (2019). The cognitive interview - a tiered approach in the real world. In J. J. Dickinson, N. S. Compo, R. N. Carol, B. L. Schwartz, & M. R. McCauley (Eds.), *Evidence-based Investigative Interviewing* (pp. 56–73). New York: Routledge.

Milne, R., Shaw, G., & Bull, R. (2007). Investigative Interviewing: The role of research. In D. Carson, R. Milne, F. Pakes, K. Shalev, & A. Shawyer (Eds.), *Applying psychology to criminal justice* (pp. 65–80). New York: John Wiley & Sons Ltd.

Monaro, M., Gamberini, L., & Sartori, G. (2017). The detection of faked identity using unexpected questions and mouse dynamics. *PLoS ONE*, *12*(5), 1–19. https://doi.org/10.1371/journal. pone.0177851

Morgan, C. A., Rabinowitz, Y. G., Hilts, D., Weller, C. E., & Coric, V. (2013). Efficacy of Modified Cognitive Interviewing, Compared to Human Judgments in Detecting Deception Related to Bio-threat Activities. *Journal of Strategic Security*, *6*(3), 100–119. https://doi.org/10.5038/1944-0472.6.3.9

Mosier, K., Fischer, U., Hoffman, R. R., & Klein, G. (2018). Expert professional judgements and "naturalistic decision making." In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge handbook of expertise and expert performance* (2nd Ed, pp. 453–475). Cambridge: Cambridge University Press.

Myklebust, T., & Bjørklund, R. A. (2006). The Effect of Long-Term Training on Police Officers' Use of Open and Closed Questions in Field Investigative Interviews of Children (FIIC). *Journal of Investigative Psychology and Offender Profiling*, *3*, 165–181. https://doi.org/10.1002/jip.52

Oberlader, V. A., Naefgen, C., Koppehele-gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of Content-Based Techniques to Distinguish True and Fabricated Statements: A Meta-Analysis. *Law and Human Behaviour*, *40*(4), 440–457. https://doi.org/10.1037/lhb0000193

Ormerod, T. C., & Dando, C. J. (2015). Finding a Needle in a Haystack: Toward a

Psychologically Informed Method for Aviation Security Screening. *Journal of Experimental Psychology: General*, *144*(1), 76–84. https://doi.org/10.1037/xge0000030

Ormerod, T. C., & Richardson, J. (2003). On the generation and evaluation of inferences from single premises. *Memory & Cognition*, *31*(3), 467–478. https://doi.org/10.3758/BF03194404

Oxburgh, G. E., & Dando, C. J. (2011). Psychology and interviewing: what direction now in our quest for reliable information? *The British Journal of Forensic Practice*, *13*(2), 135–144. https://doi.org/10.1108/14636641111134378

Oxburgh, G. E., Myklebust, T., & Grant, T. (2010). The question of question types in police interviews: a review of the literature from a psychological and linguistic perspective. *The Internation Journal of Speech, Language and the Law*, *17*(1), 45–66. https://doi.org/10.1558/ijsll.v17i1.45

Oxburgh, Gavin, Ost, J., & Cherryman, J. (2012). Police interviews with suspected child sex offenders: does use of empathy and question type influence the amount of investigation relevant information obtained? *Psychology, Crime and Law*, *18*(3), 259–273. https://doi.org/10.1080/1068316X.2010.481624

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC 2015*. Retrieved from www.LIWC.net

Perry, S. P., Murphy, M. C., & Dovidio, J. F. (2015). Modern prejudice: Subtle, but unconscious? The role of Bias Awareness in Whites' perceptions of personal and others' biases. *Journal of Experimental Social Psychology*, *61*, 64–78. https://doi.org/10.1016/j.jesp.2015.06.007

Plucker, J. A., Kaufman, J. C., Temple, J. S., & Qian, M. (2009). Do Experts and Novices Evaluate Movies the Same Way? *Psychology and Marketing*, *26*(5), 470–478. https://doi.org/10.1002/mar.20283

Porter, S., & Yuille, J. C. (1996). The Language of Deceit: An Investigation of the Verbal Clues to Deception in the Interrogation Context. *Law and Human Behaviour*, *20*(4), 443–458. https://doi.org/10.1007/BF01498980

Poyser, S., & Milne, R. (2015). No grounds for complacency and plenty for continued

vigilance: Miscarriages of justice as drivers for research on reforming the investigative interviewing process. *The Police Journal: Theory, Practice and Principles*, *88*(4), 265–280. https://doi.org/10.1177/0032258X15598951

Read, J. M., Powell, M. B., Kebbell, M. R., & Milne, R. (2009). Investigative interviewing of suspected sex offenders: a review of what constitutes best practice. *International Journal of Police Science & Management*, *11*(4), 442–459. https://doi.org/10.1350/ijps.2009.11.4.143

Reddick, S. R. (2004). Point: The case for profiling. *Internation Social Science Review*, *79*(3/4), 154–156. Retrieved from https://www.jstor.org/stable/41887190

Renoult, L., Irish, M., Moscovitch, M., & Rugg, M. D. (2019). From Knowing to Remembering: The Semantic–Episodic Distinction. *Trends in Cognitive Sciences*, *23*(12), 1041–1057. https://doi.org/10.1016/j.tics.2019.09.008

Runco, M. A., & Charles, R. E. (1993). Judgements of originality and appropriateness as predictors of creativity. *Personality and Individual Differences*, *15*(5), 537–546. https://doi.org/10.1016/0191-8869(93)90337-3

Runco, M. A., & Jaeger, G. J. (2012). The Standard Definition of Creativity. *Creativity Research Journal*, *24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Runco, M. A., Mccarthy, K. A., & Svenson, E. (1994). Judgments of the Creativity of Artwork From Students and Professional Artists. *The Journal of Psychology*, *128*(1), 23–31. https://doi.org/10.1080/00223980.1994.9712708

Shaw, D. J., Vrij, A., Leal, S., Mann, S., Hillman, J., Granhag, P. A., & Fisher, R. P. (2013). Expect the Unexpected? Variations in Question Type Elicit Cues to Deception in Joint Interviewer Contexts. *Applied Cognitive Psychology*, *27*, 336–343. https://doi.org/10.1002/acp.2911

Shih, T., Chen, C., Syu, B., & Deng, Y. (2019). A Cloud-Based Crime Reporting System with identity protection. *Symmetry*, *11*(2), 1–29. https://doi.org/10.3390/sym11020255

Sooniste, T., Granhag, P. A., Knieps, M., & Vrij, A. (2013). True and false intentions: asking about the past to detect lies about the future. *Psychology, Crime and Law*, *19*(8), 673–685. https://doi.org/10.1080/1068316X.2013.793333

Sooniste, T., Granhag, P. A., & Strömwall, L. A. (2017). Training Police Investigators to Interview to Detect False Intentions. *Journal of Police and Criminal Psychology*, *32*(2), 152–162. https://doi.org/10.1007/s11896-016-9206-9

Sooniste, T., Granhag, P. A., Strömwall, L. A., & Vrij, A. (2014). Discriminating between true and false intent among cells of suspects. *Legal and Criminological Psychology*, *21*(2), 344–357. https://doi.org/10.1111/lcrp.12063

Sooniste, T., Granhag, P. A., Strömwall, L. A., & Vrij, A. (2015). Statements about true and false intentions: Using the Cognitive Interview to magnify the differences. *Scandinavian Journal of Psychology*, *56*(4), 371–378. https://doi.org/10.1111/sjop.12216

Soukara, S., Bull, R., Vrij, A., Turner, M., & Cherryman, J. (2009). What really happens in police interviews of suspects? Tactics and confessions. *Psychology, Crime and Law*, *15*(6), 493–506. https://doi.org/10.1080/10683160802201827

Sporer, S. L. (1997). The Less Travelled Road to Truth: Verbal Cues in Deception Detection in Accounts of Fabricated and Self-Experienced Events. *Applied Cognitive Psychology*, *11*(5), 373–397. https://doi.org/10.1002/(SICI)1099-0720(199710)11:5<373::AID-ACP461>3.0.CO;2-0

Sporer, S. L. (2016). Deception and Cognitive Load: Expanding Our Horizon with a Working Memory Model. *Frontiers in Psychology*, *7*, 1–12. https://doi.org/10.3389/fpsyg.2016.00420

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A Meta-analytic synthesis. *Psychology, Public Policy, and Law*, *13*(1), 1–34. https://doi.org/10.1037/1076-8971.13.1.1

Sternberg, R. J., & Lubart, T. I. (1999). The concept of creativity: Prospects and paradigms. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 3–15). Cambridge: Cambridge University Press.

Stokoe, E. (2009). "I've got a girlfriend": Police officers doing 'self-disclosure' in their interrogations of suspects. *Narrative Inquiry*, *19*(1), 154–182. https://doi.org/10.1075/ni.19.1.09sto

Ströfer, S., Ufkes, E. G., Noordzij, M. L., & Giebels, E. (2016). Catching a Deceiver in

the Act: Processes Underlying Deception in an Interactive Interview Setting. *Applied Psychophysiology and Biofeedback*, *41*, 349–362. https://doi.org/10.1007/s10484-016-9339-8

Strömwall, L. A., & Granhag, P. A. (2003). How to Detect Deception? Arresting the Beliefs of Police Officers , Prosecutors and Judges. *Psychology, Crime and Law*, *9*, 19–36. https://doi.org/10.1080/1068316021000057659

Strömwall, L. A., Hartwig, M., & Granhag, P. A. (2006). To act truthfully: Nonverbal behaviour and strategies during a police interrogation. *Psychology, Crime and Law*, *12*(2), 207–219. https://doi.org/10.1080/10683160512331331328

Taylor, D. A., & Dando, C. J. (2018). Eyewitness Memory in Face-to-Face and Immersive Avatar-to-Avatar Contexts. *Frontiers in Psychology*, *9*, 1–11. https://doi.org/10.3389/fpsyg.2018.00507

Taylor, P. J., Dando, C., Ormerod, T. C., Ball, L. J., Jenkins, M. C., Sandham, A., & Manacere, T. (2013). Detecting insider threats through language change. *Law and Human Behaviour*, *37*(4), 267–275. https://doi.org/10.1037/lhb0000032

Truebano, M., & Munn, C. (2015). An Evaluation of the Use of Video Tutorials as Supporting Tools for Teaching Laboratory Skills in Biology. *Practice and Evidence of Scholarship of Teaching and Learning in Higher Education*, *10*(2), 121–135.

Verstijnen, I. M., van Leeuwen, C., Goldschmidt, G., Hamel, R., & Hennessey, J. M. (1998). Creative discovery in imagery and perception: Combining is relatively easy, restructuring takes a sketch. *Acta Psychologica*, *99*(2), 177–200. https://doi.org/10.1016/S0001-6918(98)00010-9

Vrij, A. (2004). Why professionals fail to catch liars and how they can improve. *Legal and Criminological Psychology*, *9*(2), 159–181. https://doi.org/10.1348/1355325041719356

Vrij, A. (2014). Interviewing to Detect Deception. *European Psychologist*, *19*(3), 184–194. https://doi.org/10.1027/1016-9040/a000201

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *TRENDS in Cognitive Science*, *10*(4), 141–142.

https://doi.org/10.1016/j.tics.2006.02.003

Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, *22*(1), 1–21. https://doi.org/10.1111/lcrp.12088

Vrij, A., Leal, S., & Fisher, R. P. (2018). Verbal Deception and the Model Statement as a Lie Detection Tool. *Frontiers in Psychology*, *9*, 1–6. https://doi.org/10.3389/fpsyt.2018.00492

Vrij, A., Leal, S., Granhag, P. A., Mann, S., Fisher, R. P., Hillman, J., & Sperry, K. (2009). Outsmarting the Liars: The Benefit of Asking Unanticipated Questions. *Law and Human Behaviour*, *33*(2), 159–166. https://doi.org/10.1007/s10979-008-9143-y

Vrij, A., Leal, S., Mann, S., & Fisher, R. (2012). Imposing cognitive load to elicit cues to deceit: inducing the reverse order technique naturally. *Psychology, Crime and Law*, *18*(6), 579–594. https://doi.org/10.1080/1068316X.2010.515987

Vrij, A., Leal, S., Mann, S., & Granhag, P. A. (2011). A Comparison between Lying about Intentions and Past Activities: Verbal Cues and Detection Accuracy. *Applied Cognitive Psychology*, *25*(2), 212–218. https://doi.org/10.1002/acp.1665

Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to Deception and Ability to Detect Lies as a Function of Police Interview Styles. *Law and Human Behaviour*, *31*(5), 499–518. https://doi.org/10.1007/s10979-006-9066-4

Vrij, A., Mann, S., Leal, S., & Fisher, R. (2012). Is anyone there? Drawings as a tool to detect deceit in occupation interviews. *Psychology, Crime and Law*, *18*(4), 377–388. https://doi.org/10.1080/1068316X.2010.498422

Walsh, D., & Bull, R. (2010). What really is effective in interviews with suspects? A study comparing interviewing skills against interviewing outcomes. *Legal and Criminological Psychology*, *15*(2), 305–321. https://doi.org/10.1348/135532509X463356

Walsh, D., & Milne, R. (2008). Keeping the PEACE? A study of investigative interviewing practices in the public sector. *Legal and Criminological Psychology*, *13*(1), 39–57. https://doi.org/10.1348/135532506X157179

Warmelink, L., Vrij, A., Mann, S., Jundi, S., & Granhag, P. A. (2012). Have you been there before? The effect of experience and question expectedness on lying about intentions. *Acta Psychologica*, *141*(2), 178–183. https://doi.org/10.1016/j.actpsy.2012.07.011

Webster, P., & Hickey, M. (1995). Challenging children to think creatively. *General Music Today*, *8*(3), 4–10. https://doi.org/10.1177/104837139500800303

Wiseman, R., Watt, C., ten Brinke, L., Porter, S., Couper, S., & Rankin, C. (2012). The Eyes Don't Have It: Lie Detection and Neuro-Linguistic Programming. *PLoS ONE*, *7*(7), 1–5. https://doi.org/10.1371/journal.pone.0040259

Appendix 1: Question Lists from Chapter 3

Question list A (Anticipated)

1. “What task did you carry out around the campus today?”
2. “How many boxes were in room A when you arrived there?”
3. “Describe the route you took from building A to building B.”
4. “Who let you in to building B?”
5. “Describe the items that you collected from building B.”
6. “How many boxes were there in the room at building B?”
7. “How difficult was the task to carry out?”
8. “Describe any discussion you had with the experimenter whilst at building B.”
9. “In relation to building B, how familiar are you with that area of the campus?”
10. “Please describe the task one final time, from start to finish. Try to be as detailed as possible.”

Question list B (Planning)

1. “What task did you carry out around the campus today?”
2. “How many boxes were in room A when you arrived there?”
3. “Describe the route you took from building A to building B.”
4. “Who let you in to building B?”
5. “Describe the items that you collected from building B.”
6. “What was the main goal of your planning?” *
7. “What was the final thing you planned?” *
8. “What was the most difficult part of your planning?” *
9. “Explain what steps you would have taken had you not been able to access building B via the main door.” **
10. “Please describe any changes you made to your plan during the planning stage.” **

* Adapted from Sooniste et al. (2013)

** Adapted from Granhag et al. (2016)

Question list C (Spatial/Temporal)

1. “What task did you carry out around the campus today?”
2. “How many boxes were in room A when you arrived there?”
3. “Describe the route you took from building A to building B.”
4. “Who let you in to building B?”
5. “Describe the items that you collected from building B.”
6. “In relation to building B, try to imagine the layout and features of the room where you collected the boxes from. Please describe this room to me, and be as detailed as you can.” *
7. “In building B, where were the boxes in relation to the door you entered through?” *
8. “How long did it take to walk from building A to building B?” *
9. “In relation to building B, other than the experimenter, where was the closest other person as you left the building?” *
10. “Please describe the task in full one last time, but now in reverse order. Try to be as detailed as possible.” **

* Adapted from Vrij et al. (2009)

** Adapted from Lancaster et al. (2013)

Appendix 2: Interviewer Questionnaire from Chapter 3

Interviewer Number:

Question List:

1. Do you think that the interviewee was telling the truth or lying?

Definitely lying						**Definitely telling the truth)**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

2. How confident are you that your judgment about whether or not the interviewee was telling the truth or lying is correct?

Not at all confident (completely unsure)						**Very confident (completely sure)**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

3. At what point did you decide whether the interviewee was telling the truth or lying?

At the beginning of the interview			**Part way through the interview**			**After the interview had ended**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

4. How difficult did you find it to decide whether the interviewee was telling the truth or lying?

Very easy						**Extremely difficult**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

5. Please explain **why** you found it easy/difficult to decide whether the interviewee was telling the truth/lying.

_____

_____

_____

_____

_____

_____

6. What type of information did you use to decide whether the interviewee was telling the truth or lying?

Only non-verbal behaviour (how the interviewee behaved)			**Both non-verbal and verbal behaviour (how the interviewee behaved and what he/she said)**			**Only verbal behaviour (what the interviewee said)**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

7. Please explain/describe the type of information you used to decide whether the interviewee was telling the truth or lying. For example, if you indicated in Q6 that you used both verbal and non-verbal behaviour, please describe/explain those behaviours in as much detail as possible.

_____

_____

_____

_____

_____

_____

_____

8. If you had to make a firm decision, would you say the interviewee was telling the truth or lying? Please tick the appropriate box.

Lying **Telling Truth**

☐ ☐

9. Please provide a further question you could have asked, that you believe is both relevant to the interview and which would **not** have been anticipated by the interviewee (please try to provide a new question each time you complete this section).

_____

_____

_____

_____

_____

10. Please write below any comments/suggestions you may have concerning your participation in this research.

_____

_____

_____

_____

_____

Thank you for participating in this study.

Appendix 3: Interviewee Questionnaire from Chapter 3

1. Age _____ 2. Gender _____

3. Occupation (if student please state whether UG or PG) _____

4. On a scale from 1 to 7 please rate **how deceptive/truthful** you were during the interview.

Totally truthful (everything I told the interviewer was truthful)			**Somewhat deceptive** (about half of what I told the interviewer was deceptive)			**Totally deceptive** (everything I told the interviewer was deceptive)
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

5. On a scale from 1 to 7 please rate **how difficult/cognitively demanding** you found the interview.

Very easy/ not at all cognitively demanding			**Somewhat difficult/ cognitively demanding**			**Very difficult/ cognitively demanding**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

6. Prior to the interview, in order to convince the interviewer that you were telling the truth about your account, to what extent did you think about what you would **say** in the interview?

I did not think about what I would say at all			**I gave some thought to what I would say**			**I thought a lot about what I would say**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

7. If you did have a strategy concerning what you would say, please describe it. Be as detailed as possible and include an explanation as to why you decided upon this strategy (later questions will be asking you about your behaviour during the interview, so this question is only about what you planned to say during the interview).

_____

_____

_____

_____

_____

_____

_____

8. If, before the interview, you did not devise a strategy concerning what you would say please explain why not**.**

_____

_____

_____

_____

_____

9. If you did devise a strategy concerning what you would say, to what extent did you actually use this strategy during the interview?

I did not use my strategy at all during the interview			**I used my strategy to some extent during the interview**			**I used my strategy to its full extent during the interview**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

10. **Prior** to the interview, in order to convince the investigator you were telling the truth, to what extent did you think about how you would **act and/or behave** during the interview?

I did not think about how I would behave at all			**I gave some thought to how I would behave**			**I thought a lot about how I would behave**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

11. If you **did** have a strategy concerning how you planned to **behave** please describe it. Be as detailed as possible and include an explanation as to why you decided upon this strategy.

_____

_____

_____

_____

_____

_____

12. If, before the interview, you did **not** devise a strategy, concerning how you would **behave,** please explain why **not**.

_____

_____

_____

_____

_____

13. If you **did** have a strategy concerning how you would **behave,** to what extent did you actually use this strategy during the interview?

I did not use my strategy at all during the interview			**I used my strategy to some extent during the interview**			**I used my strategy to its full extent during the interview**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

14. During the interview, on a scale from 1 to 7, **how motivated** were you to comply with the pre interview instructions?

Not at all motivated			**Somewhat motivated**			**Very motivated**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

15. Please write, below, further comments concerning your participation in this research.

_____

_____

_____

_____

_____

_____

_____

_____

Thank you for participating in this study.

Appendix 4: Difficulty and Anticipation Questionnaires from Chapter 3

The two following questions were repeated for each of the ten questions that the interviewee had been asked:

To what extent did you expect to be asked Question 1: "What task did you carry out around the campus today?"

Completely **Expected**			Somewhat Expected			Completely **Unexpected**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐

How difficult did you find it to answer Question 1 "What task did you carry out around the campus today?"

Completely **Expected**			Somewhat Expected			Completely **Unexpected**
1	2	3	4	5	6	7
☐	☐	☐	☐	☐	☐	☐