Applying Hierarchy of Expert Performance (HEP) to Investigative Interview Evaluation: Strengths, Challenges and Future Directions

Abstract

The purpose of this paper is to systematically examine the research literature on the decision of expert interviewers within the theoretical framework of the Hierarchy of Expert Performance (HEP, Dror, 2016). After providing an overview of the HEP framework, existing research in the investigative interviewing at each of the eight levels of the HEP framework is reviewed. The results identify areas of strength in *reliability between experts' observations* (Level 2) and of weakness in *reliability between experts' conclusions* (Level 6). Biases in investigative interview experts' decision making is also revealed at *biasability between expert conclusions* (Level 8). Moreover, no published data is available in *reliability within experts* at the level of *observations* (Level 1) or *conclusions* (Level 5), *biasability within or between expert observations* (Level 3 and 4) and *biasability within expert conclusions* (Level 7). The findings highlight areas where future research and practical endeavor are much needed investigative interview.

Keywords: Hierarchy of Expert Performance (HEP), Forensic/Investigative Interview, Bias, Professional Judgments, Decision Making.

Applying Hierarchy of Expert Performance (HEP) to Investigative Interview Evaluation: Strengths, Challenges and Future Directions

Introduction

An investigative interview, broadly defined as a structured conversation between an interviewing professional and an interviewee to gather information, is usually one of the first steps of information gathering when an alleged crime happens. The quality as well as the quantity of the information gathered plays a pivotal role in determining whether the case will move forward in the legal system. In some cases, such as child sexual abuse (CSA), the statements of alleged victims often become the primary source of evidence that the investigators rely upon (Wright & Powell, 2007). The interviewee's mind, therefore, becomes the only way to access to the crime scene. The interviewer must carefully navigate, collect and evaluate such 'psychological evidence' (i.e., interviewee's memory and accounts), using interviewing as the investigative tool.

The use of structured interview protocols (e.g., the NICHD interview protocol, Lamb, Orbach, Hershkowitz, Esplin, & Horowitz, 2007; Achieving the Best Evidence in Criminal Proceedings, Ministry of Justice, 2011) helps to obtain high-quality information from the interviewees. Some interviewers are trained experts in optimizing the interviewee's memory recall accuracy, by maximizing free-recall information from the interviewees whilst minimizing information prompted by potentially (mis)leading or suggestive questions. However, even well-trained and experienced experts can be influenced by cognitive biases (consciously or unconsciously) and make erroneous decisions (Dror, 2016; 2018). To complicate the matter, in the investigative psychology domain, the 'ground truth' is often hard to establish (for example, what actually happened), which makes evaluating expert

performance more difficult, as the accuracy cannot often be ascertained (Dror & Murrie, 2018).

The field of forensic science (e.g., DNA and fingerprint examinations) has received attention in regard to cognitive bias (e.g., in the UK, guidance by the Forensic Science Regulator, 2015; in the US, the National Academy of Forensic Science, 2009, and the President's Council of Advisors on Science and Technology, 2016). This attention is now being applied to forensic psychology (e.g., competency and sanity evaluations, Dror & Murrie, 2018; Zapf & Dror, 2017). However, research on investigative interviewing has yet to systematically paid attention to how bias may influence experts' practice and decision-making (e.g., Everson & Sandoval, 2011).

The purpose of this paper is to make a first step towards a systematic, and within a single theoretical framework, examination of the potential bias and decision making in investigative interviewing. We use the Hierarchy of Expert Performance (HEP, Dror, 2016), established in the forensic science domain (and applied to forensic psychology, Dror & Murrie, 2018), toward this aim and identify gaps in the literature as well as between knowledge and practice, while suggesting directions for future research and practical implications.

The Hierarchy of Expert Performance (HEP)

Dror (2016) suggested the following three distinct dimensions for conceptualizing, evaluating and quantifying expert performance: 1) reliability vs biasability; 2) within vs between expert performance; and 3) observations vs conclusions.

The first dimension, reliability vs biasability, are two fundamental and distinct properties of decision making (Dror, 2016; Dror & Rosenthal, 2008). *Reliability* concerns expert's performance replicability regardless of bias, such as whether different child protection case evaluators would reach the same conclusions about a case being substantiated

or not when they review the exact same investigative interviews. *Biasability* refers to the ability to make decisions based on relevant information without being biased by irrelevant contextual information (e.g., irrelevant case details which go beyond the referral question and beyond the evaluator's expertise). For instance, whether an alleged rape victim's physical attractiveness would impact the evaluation of the interviewee's credibility. There may be disagreement whether something is 'relevant' or 'irrelevant' when assessing interview credibility[1], there needs to be transparency as to what factors were considered and contributed to the decision making, regardless of what is considered relevant or not. The field of investigative interviewing needs to arrive at agreement about what factors need to be considered when evaluating the credibility of an interviewee and/or of his/her account, i.e., what is relevant, and what is irrelevant to the task of determining credibility. This is fundamental and a cornerstone underlying the profession.

The second dimension, *within* vs *between* expert, examines experts' performance compared with other experts (*between experts*), versus compared with themselves (*within experts*). For instance, a between-expert examination reviews whether different child protection case evaluators reach the same conclusions about the same interview, whereas a within-expert examination considers whether the *same* child protection case evaluator will reach the same conclusions about the same interview at different times. When an expert cannot be consistent with him/herself when making a decision based on the same exact data, this kind of unreliability cannot be explained by individual differences, thus is a more basic and problematic variability than that of between expert unreliability.

[1] Although it is not always apparent if something is relevant or not, and there are instances where it is hard to determine what is irrelevant, there are nevertheless instances that information is clearly not relevant. For example, the race of the interviewee is irrelevant and should not contribute to determining their credibility. It is not within the aim or scope of this paper to determine what is relevant, or irrelevant, but to make the point that there is some information that is irrelevant, and if it impacts the decision making, it is biasing (see discussion of this matter by the National Commission on Forensic Science (2015).

Finally, the third dimension distinguishes *observations* from *conclusions*, i.e., between how information is observed, versus how it is evaluated and conclusions are reached based on those observations. Observations are the experts' perception of the data, whereas conclusions are made from evaluating and interpreting those observations. Bias as well as reliability issues can be introduced in observations or/and conclusions, therefore we should differentiate performance that derives from the interpretation and conclusions versus that which actually derives from observations (Dror, 2016). One may misunderstand the decision if the observation from the conclusion are not teased apart. For example, forensic fingerprint examiners may reach different conclusions because they observe the same data differently in the fingermark, not because they draw conclusions differently (Dror, Champod, Langenburg, Charlton, Hunt, & Rosenthal, 2011). Therefore, it is important to distinguish as much as possible between the observation stages and the conclusions based on those observations.

An eight-level Hierarchy of Expert Performance (HEP, see Figure 1) was formulated using these three dimensions (Dror, 2016). More details and examples can be found from the forensic science and the medical domain (Dror, 2016), as well as from the forensic psychology domain (Dror & Murrie, 2018). Structuring and distinguishing between different components in expert performance allows us to have a conceptual framework to systematically examine expert performance, which helps us to understand how these elements associate with one another. In so doing gaps in the literature can be identified to formulate future research, and policies tackling particular issues can be more effectively constructed.

Applying HEP to investigative interview practice and research

Similar to forensic science and forensic psychology professionals, investigative interviewers are, in some countries, well-trained experts. They share the goal of reaching correct conclusions. Forensic scientists examine physical evidence (e.g., fingerprints, trace

materials, DNA) from the crime scenes, assess them, and write reports about them to assist investigators and court to make decisions. Investigative interviewers perform similar tasks to their forensic scientist colleagues, but they access what happened via the interviewees' minds. In many ways the nature of the information (recollections from memory, rather than physical evidence from the crime scene) makes the evaluation even more difficult, as memories are very vulnerable to many distortions, such as the effects of people's biases and 'psychological contamination' (e.g., post-event suggestions, leading and suggestive interview questions or coercive interview practices).

Although much work has been devoted to devise protocols for best practice to guide the interviewing (for reviews, see Bull, 2010; La Rooy, Brubacher, Aromäki-Stratos et al., 2015), the actual practice of the investigative interview experts has rarely been systematically examined beyond its adherence to the best-practice protocols. Using the HEP framework, we will first identify and discuss how each of the three dimensions of HEP can apply to investigative interviewing. Then we will examine what, if any, investigative interviewing research exists at each level of HEP.

Observations vs conclusions applied to investigative interviewing

Usually, it is reasonably straightforward to distinguish observations from conclusions in the forensic science domain. For instance, a fingerprint examiner will first carefully observe the minutia characteristics in the friction ridge of the fingerprints (observation), and then draw a conclusion as to whether they 'match' those on another print (Dror & Cole, 2010). In contrast, in investigative interviewing this distinction between observations and conclusions is not always well-defined, and sometimes the observations and conclusions are points along a continuum.

A more clearly-defined area of observation in interview practice is the types of questions posed by the interviewer. As demonstrated by numerous researchers (e.g., Bull,

2010; Fisher & Geiselman, 1992; Orbach & Pipe, 2011; Pipe, Lamb, Orbach, & Esplin, 2004; Vrij, Hope, & Fisher, 2014), different types of questions (such as open-ended questions, directive questions, option-posing questions or suggestive questions) used in the interview play a critical role in the accuracy of the information elicited from the interviewee. Therefore, the types of questions asked has been one of the most established areas of research and practical evaluation when examining investigative interviews. However, sometimes a question can be classified into more than one category, or difficult to categorize, thus requiring expert's judgement into the observation.

Other than the types of questions being asked, the sequence of the types of questions also plays an important role when evaluating the quality of such interviews. Several well-established guidelines (including the ABE, Ministry of Justice, 2011; Cognitive Interview, Fisher, & Geiselman, 1992; NICHD interview protocol, Lamb et al., 2007) underscore the importance of prioritising the use of open-ended questions, followed by specific questions, finally more closed questions if necessary (Bull, 2010; Newlin, Steele, Chamberlin, et al., 2015). The use of suggestive questions (questions containing information not previously mentioned by the interviewee) should be avoided, especially with child interviewees. Another related area of observation is whether the general sequence of the interview is in accordance with what is recommended in the guidelines for best practices.

Another observational factor is the information reported by the interviewees (for instance, the identity of the alleged perpetrator, the objects used for committing the alleged criminal act, the time of the alleged crime, and the location, etc.). Information provided by witnesses, victims, and suspects in interviews may be important leads in investigations, it may also elicit evidence for subsequent legal proceedings. Therefore, generally, the more information reported in the interview, the more potential leads and evidence we have to assist the investigation and legal process. However, what counts as 'investigation-relevant details'

may require the experts' subjective judgment and may vary across different kinds of investigations as well as different jurisdictions.

Interviewing is a highly dynamic interpersonal process, the interviewer's and interviewee's verbal and non-verbal behaviors can have significant impact on the interview (e.g., Almerigogna, Ost, Bull, & Akehurst, 2007; Bull & Corran, 2003; Teoh & Lamb, 2013). Therefore, the verbal and non-verbal behaviors of both interviewer and interviewee constitute another area of observation. For example, the interviewer's postures, supportiveness, whether interviewer interrupts the interviewee's response, the rapport between the interviewer and interviewee; the interviewee's body language, reluctance or refusal in answering questions, and emotions etc. could all provide valuable information about the interview. However, this is another complicated area to tackle, as the sensitivity to correctly pick up and decode others' emotional or behavioral cues in an interaction can vary according to the observers' personal characteristics, experiences and training, cultural backgrounds and many other factors. Therefore, this is an observation which requires clear and careful definitions, and at times the observer's subjective judgment may still be required.

The distinction HEP seeks to draw between observations and conclusions is important in investigative interview evaluation because it allows us to identify when and how the accuracy in expert decision making may be compromised. Conclusions are more easily identified, such as the expert/professional's judgment on the quality of the interview, the credibility of the interviewee (in countries that believe this can be determined), whether the case is substantiated, and how the case should proceed following the interview. These conclusions are often made by relevant professionals (not limited to interviewers alone, as many other professionals maybe involved in the decision-making process, such as social worker, police officer/investigator, prosecutor, judge etc.) based on their observation of the

interview. In order to assist them to make more objective judgments, structured tools for interview content analyses have been developed.

One such tool is CBCA (Criterion Based Content Analysis, CBCA, Steller & Köhnken, 1989; for recent meta-analyses, see Amado, Arce & Fariña, 2015; Amado, Arce, Fariña & Vilariño, 2016). Assessments/judgements based on CBCA are accepted as evidence in some North American courts and in criminal courts in several Western European countries such as Germany, Holland, Spain and Sweden (Steller & Böhm, 2006; Vrij, 2008), but not in the UK and Canada (Novo & Seijo, 2010). CBCA experts read written transcripts of interviews (some of which they themselves have conducted) with children in alleged sexual abuse cases and score for the presence (or absence) of 19 criteria in these transcripts. In this way, the final score is intended to help the experts to draw a more objective conclusion based on the observations (the presence of these 19 criteria). However, these criteria scores are not factual observation but require the expert raters' subjective judgement (Köhnken, 2004), therefore, we will discuss the CBCA research within the conclusion levels in the HEP. Overall, as Dror and Murrie (2018) pointed out, the forensic and investigative psychology research literature provides far more evidence in *conclusions* as compared to *observations.*

Reliability vs biasability applied to investigative interviewing

Reliability can be examined by the experts' consistency in observing the interview dynamics (observation) and making judgments about the interview (conclusions). For example, would different experts categorize the interviewer questions types in the same way if they are given the same interview recording to review (i.e., between expert observation)? Would they make the same judgement on the credibility of the interviewee (i.e., between expert conclusion)? Would the same expert make the same count of different question types if given the same interview recording to review at different times (i.e., within expert

observation)? And would s/he make the same judgement about the interview at these different times (i.e., within expert conclusions)?

Biasability, the potential impact of irrelevant contextual information (Dror, 2016), has yet to be systematically examined in investigative interview field. Interviewers and evaluators are likely to encounter a variety of contextual information that is irrelevant to their evaluation or beyond their expertise. However, the field has not yet identified what information is task-relevant or task-irrelevant in any systematic way as other forensic sciences have (see, e.g., the US National Commission on Forensic Science, 2015).

Take confirmation bias for example, the beliefs an interviewer holds about the case prior to her/his interview may, consciously or unconsciously, lead the interviewer to seek information in a certain direction (e.g., asking suggestive questions to the interviewee), thus the interview progresses in the direction that is confirming the interviewer's prior beliefs. The biases the interviewers have can be a result of (or cause) "*bias cascade"* or *"bias snowball" effect*s (Dror, 2018). With the growing knowledge and understanding in memory suggestibility, the field of investigative interviewing has identified a number of inappropriate (biased) interview practices which can lead to inaccurate reports from the interviewees, especially in children and vulnerable witnesses (for a review, see Hritz, Royer, Helm, Burd, Ojeda, & Ceci, 2015). Consequently, considerable efforts have been made to develop practical means of minimizing the possible effects of suggestibility and interviewer biases by having standard interview protocols and recommended best practices.

Although many advances have been made to address the issue of suggestibility within interview practice, much is left unexplored about how different sources of biases (e.g., see Dror, 2017) may or may not affect experts' observations as well as conclusions. In highly sensitive contexts, such as child sexual abuse cases, often with limited corroborative evidence, understanding the potential effect of biases is important, so the mitigating measures

can be developed and used. Moreover, the impact of biases can go well beyond the interview outcome and can also impact on how the case information is processed, evaluated, and finally presented at court. However, there are no standardized procedures (like Linear Sequential Unmasking (LSU) in forensic science) on deciding what information is relevant and irrelevant to present to the interviewer before conducting the interview, how the information from the interview should be presented, including what content, at which time point, and how much of it should be presented, and to whom.

Between vs within expert performance

This dimension examines experts' performance relative to other (*between*) experts versus their performance relative to themselves (*within*). For instance, examining the inter-rater reliability of assessments of the interview content is important (e.g. Lamb et al., 2007), and this addresses the level of reliability between experts' observations (Level 2 in HEP). However, examination for within experts' performance is a more fundamental and insightful measure, but it can be practically difficult to conduct, because an interviewer is more likely to recognize s/he is reviewing the same interview. Nevertheless, research on within expert performance is critical, as this is the foundation of professional practices and expertise. The unreliability within experts can severely compromise the validity of their work.

Existing research at each level of the HEP

Level 1. Reliability within expert observations

Level 1examines the most basic of expert performance: Whether the same expert, looking at the same evidence, will observe the same data. Dror et al. (2011) investigated this amongst forensic fingerprint experts and found that not only were their observations not consistent with one another (Level 2), but even the same forensic fingerprint expert looking at the same print is not always consistent with his/her own observations (Dror et al., 2011). In investigative interview research, we know of no comparable published research. However, in

Walsh's unpublished doctoral dissertation (2010) he examined real-life investigative interviews with fraud suspects (n=142 of which 115 were audio-taped and 27 verbatim transcripts). He evaluated each of these for over 30 interviewer skills, then some time later, he again evaluated every interview, but this time in a different interview order. He found that his intra-rater correlations for 25 of the 30 skills exceeded 0.90 and the lowest correlation was 0.73. These types of studies are critical for the investigative interviewing community to explore and understand potential lack of reliability within experts in observations, and take appropriate measures, when needed, to minimize them.

Level 2. Reliability between expert observations

Level 2 pertains to the reliability between experts' observations of the data. In investigative interview research, this level of examination can be demonstrated in the inter-rater reliability of the coding of interviewer question types and forensically-relevant details from interviewees. In other words, will different experts make the same observation about the types of questions posed by interviewer, and identify the same amount of relevant details in the interviewee's response? For instance, in Brown et al.'s (2017) work examining the interviewers' questions posed to vulnerable child witnesses, the trained researchers 'blind' to the research question and the researchers from the research team coded the interviewer question types (open prompts, cued invitations, direct prompts, option-posing prompts, or suggestive prompts) independently. They reached an interrater reliability of .91 (Cohen's kappa), showing a high level of reliability between their observations.

Similarly, Gagnon and Cyr (2017) reported high intra-class correlations for the number of forensically-relevant details (correlation $r = .98$) and for types of questions ($r = .99$) observed by researchers when coding the interview. When the interview question types were coded by experienced investigation professionals, the interrater agreement ranged from .57 to 1.00 (Cohen's kappa, see Walsh & Bull, 2015), and percentage of rater agreement

between police officers was 79% (Clarke & Milne, 2001; Clarke, Milne & Bull, 2011). These are results indicate a high level of reliability in between experts' observations.

Level 3. Biasability within expert observations

Level 3 in HEP examines whether the same expert will observe the evidence differently if it is presented within irrelevant contextual information. In investigative interview research, we know of no research examining this aspect of expert performance. Investigative interview experts rarely examine exactly the same case data in the way a forensic scientist might re-examine the same evidence without recognizing they were looking at a case which they had previously examined. Moreover, as discussed in previous sections on biasability and reliability, the field has not considered or identified what information is task-irrelevant or task-relevant systematically as other forensic sciences have. To minimize the possibility of experts recognizing that the interview case is one that they had examined before, researchers could use partial interview files (e.g., only one section of the interview instead of the whole interview transcript, or only viewing the interviewer questions at one occasion and viewing only the interviewee's responses at another occasion) and give different contextual information at different times.

Level 4. Biasability between expert observations

The Level 4 of HEP looks at whether the observations of the data among different experts are biased by irrelevant information. Dror (2017; see also Zapf & Dror, 2017) identified seven different sources of biasing information (see Figure 2), one of which is, the reference material (a 'target') that evidence is matched to. For example, when the evidence from the crime scene is observed within the context of the 'target suspect', will this bias how the actual evidence is observed? Although the investigative interview field has yet to systematically distinguished what information is task-irrelevant or relevant, some information (e.g., the attractiveness of an alleged rape victim) is clearly irrelevant and should not impact

experts' observations (e.g., the number of case relevant details reported in the alleged victim's interview or the types of questions asked). The authors have not found any comparable research examining biasability between-experts observation in the investigative interview literature, underscoring the need for future research.

Level 5. Reliability within expert conclusions

Level 5 is the first level in the HEP regarding experts' *conclusions* and it examines the reliability of conclusions *within* experts. This is a more basic measure of reliability, in that we would expect an expert to reach the same conclusion if considering the same data on different occasions. Findings from forensic fingerprint examiners demonstrated that they will not reach the same conclusions 10% of the time, even when the same experts examine the same pair of prints in the absence of biasing information (Ulery, Hicklin, Buscaglia & Roberts, 2012). In investigative interview research, we know of no comparable research that examines this aspect. As discussed above in Levels 1 and 3 of the HEP, *within*-expert studies are much more difficult to conduct in general, and even more so for interviewing. However, within-expert reliability regarding conclusions is critical for establishing the scientific rigor of the expertise, which should be prioritized in future research.

Level 6. Reliability between expert conclusions

Will different experts evaluating the same data reach the same conclusions in the absence of biasing information? Unfortunately, research evidence from the forensic science field showed that, even in the absence of irrelevant biasing information, fingerprint and DNA examiners will reach a range of different and conflicting conclusions when they examine the very same evidence (e.g., Dror & Hampikian, 2011; Dror & Rosenthal, 2008). Such findings are alarming as unreliable expert judgements compromise the goal of equitable justice. In investigative interview research, several studies provide information about the reliability between experts' conclusions. These can be in the form of different professionals' judgement

on credibility (e.g., is the interviewee' statement believable?), quality of interviewers' practice (e.g., is this a good interview?), whether the case is substantiated, and how the case should be proceeded.

The majority of the research evidence at this level has come from one of the most contested topics -- the credibility of child sexual abuse (CSA) testimony. Due to the difficulty of obtaining corroborative evidence in CSA cases (Herman, 2010; Lamb, Hershkowitz, Orbach, & Esplin, 2008), these investigations are often largely based on children's accounts, and the discrepant opinions about the validity and credibility of CSA cases had invoked a large amount of research since the 1990s (see Herman 2009 for reviews). However, there is still no method that can validly distinguish between truthful and false statements (Vrij, Granhag, & Porter, 2010; although several components of the CBCA have been validated, e.g., Roma, San Martini, Sabatello, Tatarelli & Ferracuti, 2011). Also, increasing research brings professionals' ability to accurately assess a child's testimony into question (Herman, 2009; Zajac, Garry, London, Goodyear-Smith, & Hayne, 2013).

A substantial body of research has demonstrated subjectivity in CSA case assessments, jeopardising the reliability of professionals' performance. Subjectivity is inferred to the degree that case ratings or determinations are unreliable, or are systematically related to characteristics of the professional, such as professional discipline (Horner et al., 1993a, 1993b; Jackson & Nuttall, 1997), years of experience (Jackson & Nuttall, 1997), the professional's gender (Finlayson & Koocher, 1991; Horner et al., 1993a, 1993b; Jackson & Nuttall, 1997), and personal history of childhood sexual abuse (Jackson & Nuttall, 1997).

Several analogue studies have focused on professionals' credibility ratings using a variety of methodologies, including observations of behavioral cues during the interview (Ceci & Crotteau-Huffman, 1997; Westcott, Davies, & Clifford, 1991), comparisons of the performance of professionals to that of laypersons (Chahal & Cassidy, 1995; Leach, Talwar,

Lee, Bala, & Lindsay, 2004; Tye, Henderson, & Honts, 1995), using verbal tools such as the CBCA (Akehurst, Koehnken, & Hoefer, 2001; Steller, Wellerhaus, & Wolf, 1988; Yuille, 1988). Across these studies and the different methods used, on average, a third of the judgements were 'incorrect' (Melkman, Hershkowitz & Zur, 2017). Overall, experts do not seem to substantially outperform laypersons in successfully identifying those telling the truth from those telling lies (Bond & DePaulo, 2006; Crossman & Lewis, 2006; Edelstein, Luten, Ekman, & Goodman, 2006; Strömwall, Granhag, & Landstrom, 2007; Vrij, Akehurst, Brown, & Mann, 2006). However, Mann, Vrij, and Bull (2004) found police performance to be noticeably above chance when observing video-recorded police interviews with real suspects; and Dando and Bull (2011) found experienced police interviewers to achieve noticeably higher accuracy when gradually disclosing information to suspects.

If we leave the issue of actual validity of the judgement (whether the professional's judgment is accurate) out, and just focus on examining the inter-rater reliability of professionals' judgment of the perceived credibility of interview contents, inter-rater reliability is still alarmingly poor, regardless of whether the statements were obtained in analogue studies (Ceci et al., 1994; Horner et al., 1993a, 1993b) or using transcripts from real investigative interviews (Finlayson & Koocher, 1991; Jackson & Nuttal, 1993). Credibility ratings of any given statement are usually distributed across the full available range in analogue studies (e.g., range from .10 to .90 in Horner et al.' studies; and 0–25% to 75–100% in Finlayson & Koocher's study) as well as field studies (ranging from "certainty that the abuse had occurred" to "certainty that it had not occurred", Finlayson & Koocher, 1991; Jackson & Nuttall, 1993).

Using structured tools, such as CBCA, to assist with analyses of the interview content seemed to improve the validity of experts' judgement on interviewee credibility (Amado et al., 2015; Roma et al., 2011). Nevertheless, the inter-rater reliability of the CBCA studies

from field data (using child sexual abuse case interview transcripts, e.g. Roma et al., 2011; Welle, Berclaz, Lacasa, Niveau, 2016) demonstrated a range of reliability coefficients for the various criteria (e.g., RE coefficient from .52 to .94 in Roma et al., 2011; Krippendorf α coefficient ranged from -.13 to .75 in Welle et al, 2016), demonstrating the need for improvement on its diagnostic accuracy, hence the unreliability of between-experts' conclusions.

In field interview data, Hershkowitz and her colleagues (Hershkowitz, Fisher, Lamb, & Horowitz, 2007) have shown that using the NICHD investigative protocol significantly improved the levels of accuracy in judgements of credibility as well as inter-rater reliability. When rating non-protocol interviews, the distribution of investigators' judgments about credibility was wider and inter-rater reliability lower (α = .764) than when rating protocol interviews (α = .874). The advantage of the protocoled interviews in increasing reliability ratings was especially evident when rating cases involving implausible allegations (α = .642 for protocol and α = .338 for non-protocol interviews), but not when rating plausible allegations (α = .811 for protocol and α = .890 for non-protocol interviews). However, rates of erroneous judgements still exceeded 40 percent, and an additional 16.7 percent (excluded from accuracy calculation) were deemed as 'no judgment possible' (Hershkowitz et al., 2007).

As for judgment on quality of the interview, when there are clear guidelines on the criteria of good practice, such as having a list of appropriate interviewer questioning skills and behavioral items (e.g. MacDonald, Snook & Milne, 2017; Read, Powell, Kebbell, Milne & Steinberg, 2014), the inter-rater reliabilities tend to be higher. For example, in MacDonald, Snook and Milne (2017)'s field evaluation of witness interview training, their inter-rater reliability (Kappa) ranged from 0.60 to 1.00 for individual items, and the overall value was 0.75. In Read et al.'s (2014) study, the agreement percentage of the behavioral categories

ranged from 50% to 100%, whereas in Walsh, King and Griffith (2017) the inter-rater correlations between two expert evaluators ranged from 0.77 to 1.0. Walsh et al. (2017) also compared the interview quality evaluation between the interviewers (who conducted the interview) and the two expert evaluators' judgements, and they found that the interviewers' self-evaluations tended to be higher than those of the experts.

Overall, similar to what Dror and Murrie (2018) found within the field of forensic psychology, most of the research literature regarding investigative interviewing is within Level 6 in the HEP *reliability between expert conclusions*. But even within this level, the data are mainly in the area of judgment of interviewee credibility and some in evaluating interview quality. Important question such as whether different experts will make the same decision about how to process the case after viewing the same interview, was unknown.

Moreover, there is a lack of reliability data for civil evaluations, and to the authors' knowledge, civil cases such as custody evaluation interviews (both of parents and children) even lack a structured protocol for their conducting (although in England civil courts require adherence to the ABE guidance). In such situations, the courts in many countries may restrict rights (or tolerate risk of great harm) based on the opinion of an evaluator. Therefore, the evaluator's ability to make reliable conclusions is critical. Data about reliability may help inform policies requiring multiple evaluations or a multi-disciplinary team (MDT) work, or to have a checklist to help measure and assess experts' judgment regarding interviews. Further understanding of the factors that affect reliability in experts' conclusions can facilitate development of specific measures and policies to improve practice, like how Fingerprint Analyses Consistency Tester (FACT; Dror et al. 2011) was developed in the forensic sciences. However, without data, we cannot ascertain the quality of the performance of the investigative interviewers, let alone characterize it, so as to develop tools to help improve it.

Level 7. Biasability within expert conclusions

Will the *same* evaluator arrive at the same conclusions when s/he reviews an identical interview within a different (irrelevant) biasing context? Evidence from forensic fingerprint experts demonstrates that, when the same fingerprints were presented to them on different occasions within different irrelevant contexts, they did not always reach the same conclusions (Dror & Rosenthal, 2008). Their conclusions varied depending on several factors, such as the difficulty of the decision, the strength of the biasing irrelevant contextual information, and the direction of the bias (for more information, see Dror, 2016).

We know of no comparable research in the investigative interview field. As discussed above (in Level 1, 3 and 5), *within*-expert studies are much more challenging to conduct. Moreover, as discussed in the section on biasability and reliability, the field of investigative interviewing has not even systematically identified what information is task-irrelevant or task-relevant. Questions, such as would an interviewer make the same judgment about the same suspect interview given two different pieces of contextual irrelevant information (e.g., in the first context the suspect had criminal history, and in the second context the suspect has no criminal record) are important to have research data on. Such studies and underlying questions, albeit more difficult to execute in the investigative interview field, are important to investigate.

Level 8. Biasability between expert conclusions

Will different experts be biased by irrelevant contextual information when examining identical materials? Several biasing effects of contextually irrelevant information have been identified in forensic science, such as whether a suspect confessed to a crime and whether other evidence suggest the suspect is the culprit. For example, research with forensic DNA experts' conclusions demonstrated that, when the examiners were exposed to the biasing information (that the suspect was involved in a gang rape), they concluded that the suspect

could not be excluded from being a contributor to the DNA mixture. In contrast, most (16 out of 17) examiners not exposed to the biasing information did not reach this conclusion (Dror & Hampikian, 2011).

In investigative interview practice, interviewers are likely to encounter a number of irrelevant contextual information before, during and after the interview. One such piece of irrelevant contextual information that is the party requesting the expert's opinion. Ideally, experts should provide impartial evidence-based opinion about the interview/case to the party who requested their services. However, research from the forensic psychology field suggests that *adversarial allegiance* (a bias towards reaching conclusions that favor the party retaining their services) may influence forensic psychology expert evaluations (Murrie, Boccaccini, Johnson, & Janke, 2008; Murrie Boccaccini, Turner, Meeks, Woods, & Tussey, 2009; Murrie & Boccaccini, 2015).

Adversarial allegiance also exists among witness suggestibility experts. MacAuliff and Arter (2016) examined 100 witness suggestibility experts' responses after they reviewed the same police interview of a female child victim, while manipulating the experts' beliefs about their retaining party (prosecution or defense) and about the interview suggestibility (low, high). They found that experts who believed they were working for the prosecution were more willing to testify than those who believed they were working for the defense when perceived interview suggestibility was low. In contrast, when perceived interview suggestibility was high, experts who believed they were working for defense were more willing to testify. Moreover, these experts' anticipated testimony focused more on pro-defense aspects of the police interview and child's memory overall (negativity bias), but favored the retaining party only when the perceived interview suggestibility was low. Additionally, the prosecution-retained experts focused more on the pro-defense aspects of the case in the high suggestibility interview condition, but focused more on the pro-prosecution

aspects in the low suggestibility interview condition, but the defense experts did not show such differences.

Another piece of information that affected the experts' judgement in MacAuliff and Arter's (2016) study was the experts' beliefs about how the interview was conducted. The experts' judgement about the child victim accuracy and interview quality were negatively affected if they believed the interview was highly-suggestive, despite the actual interview content remaining identical. However, such information (suggestibility of interview practice) is actually relevant to the quality of the interview, because it can impact the quality of information obtained from the interviewee; and these experts' judgements did rightly reflect their knowledge about it. More research is necessary in the field to identify *how* these experts make such judgements, and whether this affects their judgement at both observation as well as conclusion levels. However, the 'bottom line' is that research within the investigative interviewing domain currently does show existence of bias at the Level 8 of HEP: Different experts examining the same data can reach different conclusions due to biasing impact of irrelevant contextual information.

Information about the interviewee could also be a source of bias in experts' judgement. For instance, negative perceptions of the reliability and suggestibility of witnesses with intellectual disabilities (ID) appear to have been widespread amongst some police officers (Aarons & Powell, 2003) and legal professionals (Nathanson & Platt, 2005). In a study by Stobbs and Kebbell (2003) mock jurors judged the testimonies of adults as witnesses described as having ID as less credible, less accurate, and less competent witnesses than a typically developing (TD) counterpart, and therefore gave fewer guilty ratings of the perpetrator. Meaning that cases maybe less likely to be investigated because successful outcomes (i.e., guilty verdicts) are deemed unlikely (Aarons, Powell, & Browne, 2004).

However, some countries' recent (e.g. government) policies may be increasing the likelihood of such cases being investigated.

Peled, Iarocci, and Connolly (2004) used video and written transcripts of a child testimony to examine mock jurors' judgement of witness ability and credibility. They found that jurors believing the child had ID rated the child witness as less credible than those who believed the child was TD when the video testimony was presented; whereas such differences were not found when the testimony was presented as written transcripts. This suggested that a general (negative) bias regarding the competency of witnesses with IDs may be ameliorated when jurors are presented with actual testimony (Peled et al., 2004). Similarly, Brown and Lewis (2013) found that mock jurors judged a child witness as less accurate in responding to suggestive questions and less cognitively competent when the child witness was presented as having an ID. Such bias in perception about children with ID was also present in attorneys (Nathanson & Platt, 2005). Information about the interviewee (such as age, cultural background and disability etc.) is indeed important for the interviewer to prepare and plan the interview in order to obtain the best evidence. However, given the possible biasing effect of interviewee characteristics on evaluators' judgement, it maybe advisable to have another professional 'blind' to the interviewee characteristics to independently evaluate the credibility of the interviewee and the quality of the interview.

The field of investigative interviewing has little data on other types of task-irrelevant contextual information that may bias experts. Many questions about contextual bias are important for interviewing practice as well as policy. For example, an interviewer's workplace is technically irrelevant to an interviewee's statement credibility. But might a police interviewer from one jurisdiction reach different conclusions about the same case interview than another police interviewer from another jurisdiction (especially where actual

training differs across jurisdictions), even if they both received the same national-standardized interview protocol?

Discussion

The forensic science field had undergone a major shift in the last decade, from its previously solely object-focused inquiries (DNA, fingerprints, firearms, handwriting, etc.) to in some countries now integrating the critical role that the human experts play in forensic decision making into safeguarding best practice. Nowadays the forensic sciences are beginning to acknowledge, research, and develop regulations as well as policies, and take actions to reduce unreliability and bias in their decision making (Forensic Science Regulator, 2015; National Academy of Science, 2009; National Commission on Forensic Science, 2015; President's Council of Advisors on Science and Technology, 2016).

The investigative interview field has recognized the pivotal role that interview experts have regarding the accuracy of interviewees' statements since the 1980s, after a few day-care sex abuse scandals (such as Little Rascals (Bruck, 1998), and Wee Care (Rosenthal, 1995)). Fortunately, such cases fostered the development of evidence-based interview protocols (such as the NICHD protocol) and the introduction in England of typically video-recording investigative interviews with children and the associated 'Memorandum of Good Practice' published by the government (Home Office, 1992) to optimize interview practice and outcome.

However, the attention of the investigative interviewing field is still largely focused on the interview protocols, techniques and outcomes, with far less attention paid to the critical role that the interviewer (and other legal professionals) themselves, as human experts, play while conducting and evaluating interviews. Some fundamental issues of reliability and validity had indeed received attention, particularly in the context of interview tool

assessment. Nevertheless, much less attention has been paid to the factors affecting decision making process and the role of human experts play in investigative interview assessment.

The recent shift in forensic science has generated inspection of the factors affecting forensic science experts' performance and decision making (Dror, 2018). Research evidence in the forensic science field has yielded the Hierarchy of Expert Performance (HEP in short, Dror, 2016) model for examining expert decision making, providing a framework for systematically examining and organizing existing research to assist in identifying problematic or under-research areas. Applying HEP to investigative interviews has unveiled a few areas of strength and weakness, as well as several areas in which basic research is severely lacking. For areas of strength, investigative interview research has provided evidence on *reliability between experts' observations* (L2). These studies demonstrated reliability in the coding of interview transcripts, revealing inter-rater reliabilities among field professionals (e.g., Walsh & Bull, 2015) as well as academic researchers (e.g., Brown et al., 2017). However, research about *reliability between experts' conclusions* (L6) revealed weakness in the experts' performance reliability, even with verbal analytic tools intending to objectify and assist experts' judgement, such as CBCA. The data revealed the unreliability of between-experts' conclusions (e.g. Welle et al., 2016). However, having protocol interviews did help increase reliability (Hershkowitz et al., 2007).

As for areas lacking in research evidence, the field offers no published data on *reliability within experts* at the level of observations (L1) or conclusions (L5). This fundamental form of reliability must be examined and quantified, in order to establish the robust scientific basis for investigative interview practice. However, we acknowledge that such research can be particularly challenging to conduct, but it is imperative to establish a scientific foundation.

There is also some research at the *biasability between experts' conclusions* (L8) level, demonstrating the biasing effects of task-irrelevant information on experts' judgment (e.g., MacAuliff & Arter, 2016), though such a study requires replication. The limited available research on bias has addressed *adversarial allegiance* (MacAuliff & Arter, 2016), which may contribute to injustice in the adversarial justice system. Interviewee disability (e.g., Brown & Lewis, 2013) could also contribute to experts' biases in judgments. Biases related to gender, race, sexual orientation, attractiveness, profession and religion, or other potential topics related to crime details or criminal stereotypes, or basic base-rate expectation biases (see Figure 2), to our knowledge, still remain understudied among interview experts' decision-making. More research on biases and factors affecting investigative interview experts' decision making is clearly needed, and the field needs to thoroughly identify what information is task-relevant or task-irrelevant, and at which point of the case progression such information is actually needed.

Conclusion

Based on research in forensic science, different context management protocols could be developed to minimize such impacts of bias and irrelevant contextual information. For example, Linear Sequential Unmasking (LSU), controls the sequence and timing for providing information (Dror et al., 2015). LSU, and the need to minimize bias in forensic science, has been adopted by the UK Forensic Science Regulator (2015) as well as the US National Commission on Forensic Science (2015). However, no such research and actions appear to have been undertaken within the investigative interview field.

Countless decisions are made about any given interview, both formally and informally, at each stage of the investigative process that may influence whether a case that relies on the testimony of a witness, a victim or a suspect will proceed. For example, parents (if the interviewee is a child), social workers, police, investigative interviewers, lawyers,

expert witness, prosecutors and judges all make judgements of the capacity of a witness and the possible contribution of their evidence to a case outcome, even if the case never reaches court. The investigative interview field should develop such context management protocols based on research evidence to minimize bias in decision-making process. After all, this 'psychological crime scene' (i.e., the interviewee's memory) is highly vulnerable to contamination, resulting in consequences which can jeopardize justice. If such issues are not isolated within the judicial process, they can easily cause bias cascade or bias snowball that stand in the way of the fair administration of justice.

Author Statement: The authors declare no conflict of interest for this manuscript's submission, and all authors have approved the manuscript for its submission. This article does not contain any studies with human participants or animals performed by any of the authors. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Aarons, N. M., & Powell, M. B. (2003). Issues related to the interviewer's ability to elicit reports of abuse from children with an intellectual disability: A review. *Current Issues in Criminal Justice, 14*, 257–268. Retrieved from http://search.informit.com.au/fullText;dn=20032246;res=AGISPT

Aarons, N. M., Powell, M. B., & Browne, J. (2004). Police perceptions of interviews involving children with intellectual disabilities: A qualitative inquiry. *Policing and Society, 14,* 269–278. doi:10.1080/1043946042000241848

Akehurst, L., Koehnken, G., & Hoefer, E. (2001). Content credibility of accounts derived from live and video presentations. *Legal and Criminological Psychology, 6,* 65–83.

Almerigogna, J., Ost, J., Bull, R., & Akehurst, L. (2007). A state of high anxiety: How unsupportive interviewers can increase the suggestibility of child witnesses. *Applied Cognitive Psychology, 21,* 963–974.

Amado, B. G., Arce, R., & Fariña, F. (2015). Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context*, *7*(1), 3-12.

Amado, B. G., Arce, R., Fariña, F. & Vilariño, M. (2016) Criteria-Based Content Analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology, 16*, 201-210.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10(3),* 214–234.

Brown, D. A. & Lewis, C. N. (2013) Competence is in the Eye of the Beholder: Perceptions of intellectually disabled child witnesses, *International Journal of Disability, Development and Education, 60(1),* 3-17. doi:10.1080/1034912X.2013.757132

Brown, D., Lewis, C., Stephens, E. & Lamb, M.E. (2017) Interviewers' approaches to questioning vulnerable child witnesses: The influences of developmental level versus intellectual disability status. *Legal and Criminological Psychology, 22,* 332–349.

Bruck, M. (1998). The trials and tribulations of a novice expert witness. In S. J. Ceci & H. Hembrooke (Eds.), Expert witnesses in child abuse case: What can and should be said in court (pp. 85–104). Washington, DC: American Psychological Association.

Bull, R. (2010). The investigative interviewing of children and other vulnerable witnesses: Psychological research and working/professional practice. *Legal and Criminological Psychology, 15(1),* 5-23.

Bull, R., & Corran, E. (2003). Interviewing child witnesses: Past and future. *International Journal of Police Science and Management, 4,* 315–322.

Ceci, S. J., & Crotteau-Huffman, M. L. (1997). How suggestible are preschool children? Cognitive and social factors. *Journal of the American Academy of Child and Adolescent Psychiatry, 36,* 948–958.

Ceci, S. J., Loftus, E. F., Leichtman, M. D., & Bruck, M. (1994). The possible role of source misattributions in the creation of false beliefs among preschoolers. *International Journal of Clinical and Experimental Hypnosis, 42,* 304–320.

Chahal, K., & Cassidy, T. (1995). Deception and its detection in children: A study of adult accuracy. *Psychology, Crime & Law, 1,* 237–245. doi:10.1080/10683169508411959

Clarke, C., & Milne, R. (2001). National evaluation of the PEACE investigative interviewing course. Police Research Award Scheme. London: Home Office

Clarke, C., Milne, B., & 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. http://dx.doi.org/10.1002/jip.144

Crossman, A. M., & Lewis, M. (2006). Adults' ability to detect children's lying. *Behavioral Sciences & the Law, 24(5)*, 703-715.

Dando, C. J., & Bull, R. (2011). Maximising opportunities to detect verbal deception: Training police officers to interview tactically. Journal of Investigative Psychology and Offender Profiling, 8(2), 189-202. doi: 10.1002/jip.145

Dror, I. E. (2016). A Hierarchy of Expert Performance. *Journal of Applied Research in Memory and Cognition, 5 (2)*, 121-127.

Dror, I. E. (2017). Human expert performance in forensic decision making: Seven Different Sources of Bias. *Australian Journal of Forensic Sciences, 49 (5),* 541-547.

Dror, I. E. (2018). Biases in Forensic Experts. *Science, 360 (6386)*, 243. doi: 10.1126/science.aat8443

Dror, I. E., Champod, C., Langenburg, G., Charlton, D., Hunt, H., & Rosenthal, R. (2011). Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a 'target' comparison. *Forensic Science International, 208(1–3)*, 10 –17. doi:10.1016/j.forsciint.2010.10.013

Dror, I. E., & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. *Science & Justice, 51*, 204–208.

Dror, I. E., & Murrie, D. C. (2018). A Hierarchy of Expert Performance Applied to Forensic Psychological Assessments. *Psychology, Public Policy, and Law, 24 (1),* 11-23. doi:10.1037/law0000140

Dror, I. E., & Rosenthal, R. (2008). Meta-analytically quantifying the reliability and biasability of forensic experts. *Journal of Forensic Sciences, 53*, 900–903. doi: 10.1111/j.1556-4029.2008.00762.x

Dror, I. E., Thompson, W. C., Meissner, C. A., Kornfield, I., Krane, D., Saks, M., & Risinger, M. (2015). Context management toolbox: A linear sequential unmasking

(LSU) approach for minimizing cognitive bias in forensic decision making. *Journal of Forensic Sciences, 60,* 1111–1112.

Earwaker, H., Morgan, R. M., Harris, A. J. L., & Hall, L. J. (2015). Fingermark submission decision-making within a UK fingerprint laboratory: Do experts get the marks that they need? *Science & Justice: Journal of the Forensic Science Society, 55,* 239–247. doi:10.1016/j.scijus.2015.01.007

Edelstein, R. S., Luten, T. L., Ekman, P., & Goodman, G. S. (2006). Detecting lies in children and adults. *Law and Human Behavior, 30(1),* 1-10.

Everson, M. D. & Sandoval, J. M. (2011) Forensic child sexual abuse evaluations: Assessing subjectivity and bias in professional judgements. *Child Abuse & Neglect, 35,* 287-298. doi:10.1016/j.chiabu.2011.01.001

Finlayson, L. M., & Koocher, G. P. (1991). Professional judgement and child abuse reporting in sexual abuse cases. *Professional Psychology Research Practice,22(6),* 464–472. doi: 10.1037/0735-7028.22.6.464

Fisher, R. P., & Geiselman, R. E. (1992). *Memory enhancing techniques for investigative interviewing: The cognitive interview.* Springfield, IL: Charles C Thomas.

Forensic Science Regulator (2015). Cognitive bias effects relevant to forensic science examinations: Guidance. Birmingham: The Forensic Science Regulator. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/510147/217_FSR-G-217_Cognitive_bias_appendix.pdf

Gagnon, K. & Cyr, M. (2017) Sexual abuse and preschoolers: Forensic details in regard of question types. *Child Abuse & Neglect, 67*, 109–118. doi: 10.1016/j.chiabu.2017.02.022

Herman, S. (2009). Forensic child sexual abuse evaluations: Accuracy, ethics, and admissibility. In K. Kuehnle, & M. Connell (Eds.), The evaluation of childsexual

abuse allegations: A comprehensive guide to assessment and testimony (pp. 247–266). Hoboken, NJ: John Wiley and Sons.

Herman, S. (2010). The role of corroborative evidence in child sexual abuse evaluations. *Journal of Investigative Psychology and Offender Profiling, 7(3),*189–212. doi: 10.1002/jip.122

Hershkowitz, I., Fisher, S., Lamb, M. E., & Horowitz, D. (2007). Improving credibility assessment in child sexual abuse allegations: The role of the NICHD investigative interview protocol. *Child Abuse & Neglect, 31*, 99–110. doi:10.1016/j.chiabu.2006.09.005

Home Office. (1992) Memorandum of good practice on video recorded interviews with children for criminal proceedings. London. Her Majesty's Stationery Office.

Horner, T. M., Guyer, M. J., & Kalter, N. M. (1993a). Clinical expertise and the assessment of child sexual abuse. *Journal of the American Academy of Child and Adolescent Psychiatry, 32*, 925–931.

Horner, T. M., Guyer, M. J., & Kalter, N. M. (1993b). The biases of child sexual abuse experts: Believing is seeing. *Bulletin of the American Academy of Psychiatry and the Law, 21*, 281–292.

Hritz, A. C., Royer, C. E., Helm, R. K., Burd, K. A., Ojeda, K., & Ceci, S. J. (2015). Children's suggestibility research: Things to know before interviewing a child. Anuario de Psicología Jurídica, 25(1), 3-12. doi: 10.1016/j.apj.2014.09.002

Jackson, J. L. (1996). Truth or fantasy: The ability of barristers and laypersons to detect deception in children's testimony. Paper presented at the AP-LS Biennial Conference, Hilton Head Island, South Carolina.

Jackson, H., & Nuttall, R. (1993). Clinician responses to sexual abuse allegations. *Child Abuse & Neglect, 17(1)*, 127–143. doi:10.1016/0145-2134(93)90013-U

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition, 2 (1)*, 42-52.

Köhnken, G. (2004). Statement Validity Analysis and the detection of the truth. In P. A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (pp. 41-63). Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511490071.0

Lamb, M.E., Hershkowitz, I., Orbach, Y. and Esplin, P.W. (2008). Tell Me What Happened, Wiley, Chichester.

Lamb, M.E., Orbach, Y., Hershkowitz, I., Esplin, P.W., & Horowitz, D. (2007). Structured forensic interview protocols improve the quality and effectiveness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol. *Child Abuse & Neglect 31(11–12):*1201–1231.

La Rooy, D., Brubacher, S. P., Aromäki-Stratos, A., Cyr, M., Hershkowitz, I., Korkman, J., Myklebust, T., Naka, M., Peixoto, C. E., Robertsj K. P., Stewart, H., & Lamb, M. E. (2015). The NICHD Protocol: A review of an internationally-used evidence-based tool for training child forensic interviewers. *Journal of Criminological Research, Policy and Practice, 2*, 76 – 89.

Leach, A. M., Talwar, V., Lee, K., Bala, N., & Lindsay, R. C. L. (2004). Intuitive" lie detection of children's deception by law enforcement officials anduniversity students. *Law and Human Behavior, 28(6),* 661–685. doi:10.1007/s10979-004-0793-0

MacDonald, S. & Snook, B. & Milne, R. (2017) Witness Interview Training: a Field Evaluation. *Journal of Police and Criminal Psychology, 32*, 77–84. doi:10.1007/s11896-016-9197-6

Mann, S., Vrij, A., & Bull, R. (2004). Detecting true lies: police officers' ability to detect suspects' lies. *Journal of Applied Psychology, 89(1)*, 137-149.

McAuliff, B. D., & Arter, J. L. (2016) Adversarial Allegiance: The Devil is in the Evidence Details, Not Just on the Witness Stand. *Law and Human Behavior, 40(5),* 524-535. doi:10.1037/lhb0000198

Melkman, E. P., Hershkowitz, I., & Zur, R. (2017). Credibility assessment in child sexual abuse investigations: A descriptive analysis. *Child Abuse & Neglect, 67*, 76-85.

Ministry of Justice (2011), Achieving Best Evidence in Criminal Proceedings: Guidance on Interviewing Victims and Witnesses, and Guidance on Using Special Measures, Ministry of Justice, London.

Murrie, D. C., & Boccaccini, M. T. (2015). Adversarial allegiance among expert witnesses. *Annual Review of Law and Social Science, 11*, 37–55. doi: 10.1146/annurev-lawsocsci-120814-121714

Murrie, D. C., Boccaccini, M. T., Turner, D. B., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law, 15,* 19–53. doi:10.1037/a0014897

Murrie, D. C., Boccaccini, M. T., Zapf, P. A., Warren, J. I., & Henderson, C. E. (2008). Clinician variation in findings of competence to stand trial. *Psychology, Public Policy, and Law, 14*, 177–193. doi:10.1037/a0013578

Nathanson, R., & Platt, M. D. (2005). Attorneys' perceptions of child witnesses with mental retardation. *The Journal of Psychiatry and Law, 33*, 5–42. Retrieved from http://search.proquest.com/docview/620844481?accountid=14782

National Academy of Science (2009). Strengthening forensic science in the United States: A path forward. Washington, DC: The National Academies Press. https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf

National Commission on Forensic Science (2015). Ensuring that forensic analysis is based upon task-relevant information. https://www.justice.gov/ncfs/file/818196/download

Newlin, C., Steele, L. C., Chamberlin, A., Anderson, J., Kenniston, J., Russell, A., Stewart, H. & Vaughan-Eden, V. (2015) *Child forensic interviewing: Best practices*, pp. 1–20.

Orbach, Y. & Pipe, M. E. (2011) Investigating substantive issues, in Lamb, M.E., La Rooy, D.J., Malloy, L.C. and Katz, C. (Eds), *Children's Testimony: A Handbook of Psychological Research and Forensic Practice*, Wiley, Chichester, pp. 147-63.

Peled, M., Iarocci, G., & Connolly, D. A. (2004). Eyewitness testimony and perceived credibility of youth with mild intellectual disability. *Journal of Intellectual Disability Research, 48*, 699–703. doi:10.1111/j.1365-2788.2003.00559.x

Pipe, M. E., Lamb, M. E., Orbach, Y., & Esplin, P. W. (2004). Recent research on children's testimony about experienced and witnessed events. *Developmental Review, 24,* 440–468. http://dx.doi.org/10.1016/j.dr.2004.08.006

President's Council of Advisors on Science and Technology (2016). Report to the President: Forensic science in the criminal courts. Washington, DC: Executive Office of the President of the United States.

Read, J., Powell, M., Kebbell, M., Milne, B., & Steinberg, R. (2014). Evaluating police interviewing practices with suspects in child-sexual abuse cases. *Policing and Society, 24(5),* 523-544.

Roma P, Martini PS, Sabatello U, Tatarelli R, Ferracuti S. (2011) Validity of criteria based content analysis (CBCA) at trial in free-narrative interviews. *Child Abuse Neglect, 35,* 613-620.

Rosenthal, R. (1995). State of New Jersey v Margaret Kelly Michaels: An overview. *Psychology, Public Policy, and Law, 1,* 246–271.

Steller, M., & Böhm, C. (2006). Cincuenta años de jurisprudencia del Tribunal Federal Supremo alemán sobre la psicología del testimonio. Balance y perspectiva [Fifty years of the German Federal Court jurisprudence on forensic psychology]. In T. Fabian, C. Böhm, & J. Romero (Eds.), *Nuevos caminos y conceptos en la psicología jurídica* (pp. 53-67). Münster, Germany: LIT Verlag.

Steller, M., & Köhnken, G. (1989). Criteria-based content analysis. In D. C. Raskin (Ed.), Psychological methods in criminal investigation and evidence (pp. 217-245). New York, NY: Springer-Verlag.

Steller, M., Wellerhaus, P., & Wolf, T. (1988, June). *Empirical validation of criteria based content analysis.* Paper presented at NATO Advanced Study Institute on Credibility Assessment, Maratea, Italy.

Stobbs, G., & Kebbell, M. R. (2003). Jurors' perception of witnesses with intellectual disabilities and the influence of expert evidence. *Journal of Applied Research in Intellectual Disabilities, 16*, 107–114. doi:10.1046/j.1468-3148.2003.00151.x

Strömwall, L. A., Granhag, P. A., & Hartwig, M. (2004). Practitioners' beliefs about deception. In P. A. Granhag & L. A. Strömwall (Eds.), Deception detection in forensic contexts (pp. 229-250). Cambridge, UK: Cambridge University Press.

Teoh, Y.S., & Lamb, M. E. (2013). Interviewer demeanor in forensic interviews of children. *Psychology, Crime & Law, 19(2)*, 145–159.

Tye, M. J. C., Henderson, S. A., & Honts, C. R. (1995, January). Evaluating children's testimonies: CBCA and lay subjects. Little Rock, AR: Paper presented at meeting of CRIME CON: International Internet Conference on Crime and Criminal Justice.

Vrij, A. (2008). Detecting lies and deceit: Pitfalls and opportunities (2nd ed.). Chichester, UK: John Wiley.

Vrij, A., Akehurst, L., Brown, L., & Mann, S. (2006). Detecting lies in young children, adolescents and adults. *Applied Cognitive Psychology, 20(9),* 1225–1237. doi:10.1002/acp.1278

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest, 11,* 89–121. doi: 10.1177/1529100610390861

Vrij, A., Hope, L., & Fisher, R. P. (2014). Eliciting reliable information in investigative interviews. *Policy Insights from the Behavioral and Brain Sciences, 1(1),* 129-136. Doi: 10.1177/2372732214548592

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS ONE, 7(3)*, e32800. doi: 10.1371/journal.pone.0032800

Walsh, D. (2010). Towards a framework for interviewing suspects of fraud. Doctoral dissertation. Leister University.

Walsh, D. & Bull, R. (2015). Interviewing suspects: examining the association between skills, questioning, evidence disclosure, and interview outcomes. *Psychology, Crime & Law, 21 (7)*, 661-680.

Walsh, D., King, M., & Griffiths, A. (2017). Evaluating interviews which search for the truth with suspects: but are investigators' self-assessments of their own skills truthful ones? *Psychology, Crime & Law, 23 (7),* 647-665. doi: 10.1080/1068316X.2017.1296149

Weed, L. L. (1970). *Medical records, medical evaluation, and patient care: The problem-oriented medical record as a basic tool*. Cleveland, OH: Press of Case Western Reserve University.

Welle, I., Berclaz, M., Lacasa, M. J., & Niveau, G. (2016). A call to improve the validity of criterion-based content analysis (CBCA): Results from a field-based study including 60 children's statements of sexual abuse. *Journal of forensic and legal medicine, 43,* 111-119.

Westcott, H. L., Davies, G. M., & Clifford, B. R. (1991). Adults' perceptions of children's videotaped truthful and deceptive statements. *Children & Society, 5*,123–135.

Wright, R., & Powell, M. B. (2007). What makes a good investigative interviewer of children?: A comparison of police officers' and experts' perceptions. *Policing: An International Journal of Police Strategies and Management, 30*, 21-31.

Yuille, J. C. (1988). The systematic assessment of children's testimony. *Canadian Psychologist, 29*, 247–262.

Zajac, R., Garry, M., London, K., Goodyear-Smith, F., & Hayne, H. (2013). Misconceptions about childhood sexual abuse and child witnesses: Implications for psychological experts in the courtroom. *Memory, 21(5),* 608–617. doi: 10.1080/09658211.2013.778287

Zapf, P. A., & Dror, I. E. (2017). Understanding and mitigating bias in forensic evaluation: Lessons from forensic science. *International Journal of Forensic Mental Health, 16 (3)*, 227-238. doi:10.1080/14999013.2017.1317302

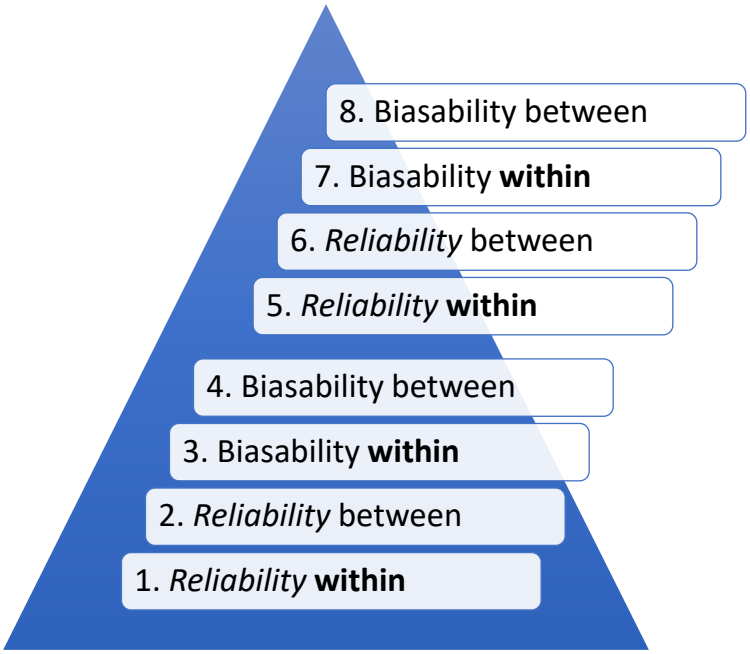

Figure 1: HEP: Hierarchy of Expert Performance (Dror, 2016)

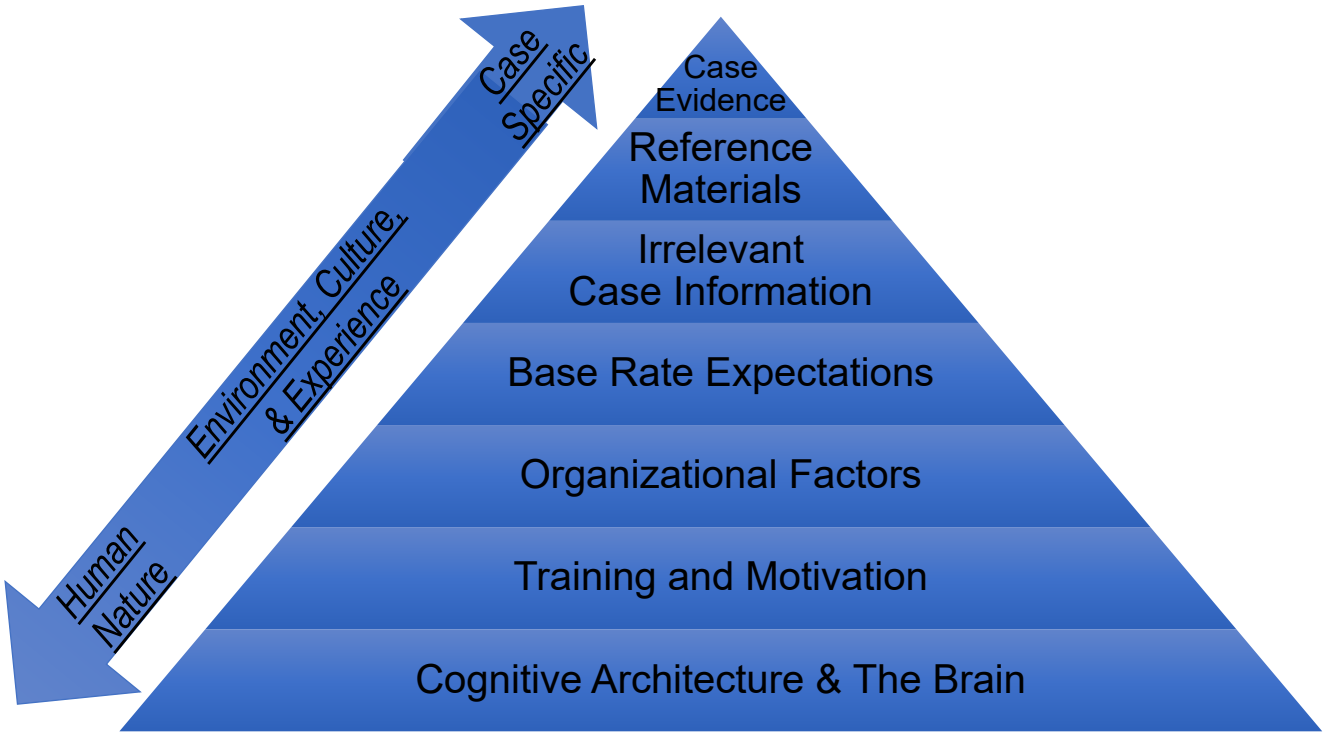

Figure 2: Seven different sources of biasing information

(Dror, 2017; Zapf & Dror, 2017).