Journal of Applied Research in Memory and Cognition 5 (2016) 121–127

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

# A Hierarchy of Expert Performance☆

Itiel E. Dror*

University College London, United Kingdom

Expert performance can be quantified by examining reliability and biasability between and within experts, and teasing apart their observations from their conclusions. I utilize these parameters to develop a Hierarchy of Expert Performance (HEP) that includes eight distinct levels. Using this hierarchy I evaluate and quantify the performance of forensic experts, a highly specialized domain that plays a critical role in the criminal justice system. Evaluating expert performance within HEP enables the identification of weaknesses in expert performance, and enables the comparison of experts across domains. HEP also provides theoretical and applied insights into expertise.

The successful completion of many everyday tasks depends upon human performance abilities; even more so in highly specialized domains, such as aviation, medicine, forensics, finance, and policing, where expert performance is critical. In these and other domains, we trust our money, health, and even our lives to the hands of experts. Most experts are talented people with many years of training and experience and therefore are in high demand and well paid.

But, how well do competent experts actually perform? Even defining expertise and who is an expert has been a complex and challenging task with a variety of views and disagreements (Feldon, 2007; Hoffman, 1996). Advances in quantifying expert performance are mostly limited to domain-specific areas of expertise, and hence there is a lack of measurements and a unified framework to assess expert performance across domains.

An important component of expert performance across domains relates to biasability and reliability. Biasability refers to the ability to make decisions based on relevant information without being biased by irrelevant contextual information. For example, a forensic expert who is aware of such information (e.g., that the suspect confessed to the crime, that eyewitnesses identified the suspect, or that the detective believes the suspect is guilty) is biased to incorrectly judge that the forensic evidence (e.g., firearms, handwriting, voice, fingerprints, etc.) matches the suspect and can wrongly identify the suspect (Dror & Cole, 2010).

In addition to issues of biasability there are basic reliability issues. That is, how reliable (i.e., how consistent, how reproducible) is expert decision making even when there is no exposure to irrelevant biasing information? Would experts examining the same forensic evidence reach the same conclusions (e.g., identify the same suspect)? Reliability issues are even more fundamental than biasability, as biasability builds on reliability, adding irrelevant contextual information. Reliability examines the consistency of expert performance even without such biasing influences. Therefore, it touches upon a more basic element of expert performance.

Biasability and reliability can be examined *between* experts. For example, reliability between experts would quantify the extent to which two medical radiologists would reach a similar diagnosis when examining the same X-ray. While individual differences exist in everyday decision making, when it comes to experts we expect that such differences are minimal (if they exist at all). We expect that experts reach the same conclusions or, at the very least, similar conclusions. If two experts reach different, conflicting, decisions then we wonder who got it wrong. This is especially acute in the more scientific and objective expert domains.

At a more basic level we can examine biasability and reliability *within* experts. Here, rather than comparing performance across different experts, biasability and reliability are compared for the same expert at different times. That is, reliability within experts quantifies the extent to which the same expert reaches similar conclusions when examining the same data at different times. While the lack of reliability between experts is concerning, the lack of reliability within experts is alarming. Expert decision-making should be driven by data and expertise; therefore, the same expert, looking at an identical situation, should generally reach the same conclusion.

☆ Please note that this paper was handled by the current editorial team of JARMAC.

* Correspondence concerning this article should be addressed to Itiel E. Dror. Contact: i.dror@ucl.ac.uk.

Biasabilty within experts quantifies the extent to which the same expert would reach similar conclusions when examining the same data presented within irrelevant contextually biasing information. Here the measurement is whether the expert was biased by the irrelevant contextual information. In contrast to pure reliability, where the identical case is presented twice, here contextually irrelevant information is added.

Another important distinction in expert performance is between the information observed and the conclusions made on the basis of those observations. For example, a medical doctor may observe that a patient is pale, disoriented and has a blood pressure of 90/60, then, based on these observations, determine the problem and course of action.

This distinction is important, but these elements are often lumped together. For instance, risk taking decisions are comprised from two distinct elements: first, the perception of risk and, second, what to do given the level of perceived risk. If these two elements are lumped together and not unpacked then risk taking decisions might appear to underpin a behaviour when, in fact, the behaviour might result from a lack in the perception of risk rather than a risk-taking decision.

In the medical domain the distinction between the 'observational' and the 'conclusion' elements has been formally separated with the use of SBAR (Thomas, Bertram, & Johnson, 2009; Wacogne & Diwakar, 2010). SBAR is used to facilitate communication and stands for Situation, Background, Assessment, and Recommendation. The first half, Situation and Background, provide observations, whereas the second half, Assessment and Recommendation, provide conclusions.

Given the parameters described above (biasability vs. reliability, within vs. between, observations vs. conclusions) and their hierarchical nature (reliability is more basic than biasability; within-expert performance is more basic than between-expert performance; and conclusions are based on observations), expert performance can be organized in a clear 8-level hierarchy. Below, I describe this Hierarchy of Expert Performance (HEP) and use forensic experts – a highly specialized domain wherein important scientific evaluations are made – to illustrate the different levels of HEP.

At the very top of HEP (Level 8) is the between-expert biasability of conclusions. The measurement of expert performance at this level examines whether different experts, making decisions on identical data, would be biased by irrelevant contextual information. For example, would forensic DNA experts be biased by the details of the case? This was one of the research questions examined by Dror and Hampikian (2011). In a DNA case (from an actual adjudicated criminal case) one of the assailants in a gang rape testified against another suspect as part of a plea bargain. Evidence from the crime scene was examined by DNA experts, who were aware of the testimony against the suspect. Furthermore, additional potentially biasing information was the fact that without corroborating evidence from the DNA experts the plea bargain testimony would not be admissible in court. The DNA evidence from the crime scene was a mixture (composed from multiple contributors) and therefore required some judgement and interpretation. With the biasing irrelevant contextual information, the two experts who examined the DNA mixture both concluded that the suspect in question could not be excluded from being a contributor to the DNA mixture from the crime scene.

To test whether the irrelevant contextual information biased their conclusion, Dror and Hampikian (2011) took the original DNA evidence and presented it to 17 experts, but did not include the irrelevant contextual information to which both of the original experts were exposed. The findings showed that only 1 out of the 17 experts agreed with the experts who were exposed to the biasing information. Computing the exact *p*-value and the associated effect size (*r*-equivalent) based on the Fisher Exact Test, gives a *p*-value of .018 (one tail) and an *r*-equivalent effect size of .49; therefore, Dror and Hampikian (2011) concluded that the data suggested that "the extraneous context of the criminal case may have influenced the interpretation of the DNA evidence" (p. 204).

Murrie, Boccaccini, Guarnera, and Rufino (2013) manipulated contextually irrelevant information by providing the same evidence to different experts, but manipulating the experts belief for whom they were working for (they told some of the experts that they were hired by the defence, while other experts were told that they were hired by the prosecution). This irrelevant context biased the experts' conclusions for three of the four cases presented (Cohen's *d* for the three cases with significant effects ranged from 0.55 to 0.85). Hence, again, research shows the biasing effect of irrelevant contextual information (for a discussion of what is considered irrelevant, see the National Commission on Forensic Science, 2015).

Level 8 in the Hierarchy of Expert Performance (HEP) addresses such biasing effects *between* experts. Level 7 examines such biases *within* experts. Level 7 is under Level 8 because it examines a more basic question: Would the *same* expert reach the same (or a different) conclusion when an identical case is presented within a different, irrelevant, biasing context. To investigate this question, fingerprint experts were presented with cases that they had examined in the past. Without their knowledge, these cases were re-presented to them; however, irrelevant contextual information was manipulated. For example, if in the past the experts concluded that the fingerprints matched, the fingerprints were now presented within the context that 'someone else confessed to the crime,' or that 'the suspect has a solid alibi.' These manipulations were used in two studies (Dror & Charlton, 2006; Dror, Charlton, & Peron, 2006), and their meta-analysis is presented in Dror and Rosenthal (2008).

The findings demonstrated that fingerprint experts could be biased by irrelevant contextual information. The within-expert data showed that experts changed their conclusions between 17% and 80% of the time when the same data were presented for a second time within contextually-irrelevant biasing information (for a review of findings and proposed solutions see Dror & Cole, 2010; Kassin, Dror, & Kukucka, 2013). The presence and strength of the biasing effects on the experts' conclusions was dependent upon three factors:

1. *The strength of the biasing information*. Some irrelevant information is more biasing than other information. For example, a solid alibi has a stronger impact than 'the detective does

not believe this suspect is guilty.' Many forensic examiners believe that they are objective and immune from the effects of contextually biasing information (e.g., "examiners who want to read investigative reports or talk to investigators before or while they examine a case. Perhaps such interest merely provides some personal satisfaction which allows them to enjoy their jobs *without actually altering their judgment*. Admittedly, many tasks performed by forensic examiners can be tedious, mundane, and to many people just not interesting or diverse enough to interest them in doing it" (Butt, 2013, p. 60 emphasis added)). Recently there is a growing acknowledgment that irrelevant contextual information can bias forensic conclusions and there is a move to require forensic experts to avoid such irrelevant biasing information (see coverage by *Science* (Servick, 2015) and by *Nature* (Spinney, 2010); as well as the US National Commission on Forensic Science document on "Ensuring that forensic analysis is based upon task-relevant information," 2015; and the UK Forensic Regulator guidance document on "Cognitive bias effects relevant to forensic science examinations," 2015).

2. *The difficulty of the decision*. When the conclusion is easy, clear-cut, then the biasing information does not impact the final conclusion as much as when it is difficult, near the decision threshold. When the evidence from the crime scene is low in quantity and in quality, there is more leeway for irrelevant contextual information to influence expert conclusions. The strength of biasing information, along with the difficulty of the decision, combined, determine the 'bias danger zone' – those instances where bias is most likely to impact the forensic conclusion.
3. *The direction of the bias*. It is easier to bias forensic experts towards the non-committal conclusion of 'inconclusive' than to the definitive 'identification' conclusion (which they may also need to defend in court). Nevertheless, the data also show that previous highly confident 'identification' conclusions can be changed to 'inconclusive' or 'exclusion' when the same fingerprints were presented to the same experts within irrelevant contextual information.

Level 6 of HEP does not examine biasability, it quantifies the more basic performance measure of reliability. That is, even without biasing irrelevant contextual information, are experts consistent with one another? A number of studies with forensic DNA experts have examined this question and have revealed an alarming lack of reliability (when dealing with complex DNA, such as mixtures with multiple contributors, a great deal is determined at the discretion of the forensic DNA expert (Starr, 2016)). In the first published study examining this question, Dror and Hampikian (2011) presented 17 experts with an identical DNA mixture. Although the 17 experts were from the same lab, and followed the same procedures and protocols, their conclusions varied: Twelve concluded that the suspect could not have contributed to the DNA mixture, four concluded that it was inconclusive, and one concluded that the suspect could have contributed to the DNA mixture.

A study conducted by the National Institute of Standards and Technology (Coble, 2015) replicated this finding using DNA that could be analysed with established statistical tools (used regularly in court). Similar to Dror and Hampikian (2011), Coble (2015) found significant differences and variations in the forensic conclusions, even within the same forensic laboratories, using the same statistics (Starr, 2016). Similar findings have also been observed in other expert forensic domains, such as footwear identification conclusions (Majamaa & Ytti, 1996).

Level 5 in HEP examines the reliability of conclusions *within* experts. Rather than measuring differences *between* experts (Level 6), the level underneath it quantifies a more basic measure, the reliability *within* the same expert. In fingerprinting, Dror and Charlton (2006; see also Dror & Rosenthal, 2008) found that the same fingerprint examiner would reach a different conclusion when the same evidence was presented 8% of the time (even without manipulating context/biasability).

A study conducted by the Federal Bureau of Investigation (Ulery, Hicklin, Buscaglia, & Roberts, 2012) replicated the initial findings by Dror and his colleagues. In the FBI study, 72 expert forensic fingerprint examiners compared the same 25 pairs of fingerprints on two different occasions, approximately seven months apart. Their data showed that fingerprint experts were inconsistent with their own conclusions on the same pair of fingerprints 10% of the time (Ulery et al., 2012).

Levels 5–8 in HEP, discussed thus far, pertain to experts' *conclusions*; however, these conclusions are based on observations of the data and evidence. For example, fingerprint experts compare the minutia (fingerprint characteristics – see Figure 2) that they observe on a latent fingerprint from the crime scene with those from the fingerprint of a suspect. They assess whether these are similar enough to conclude that both fingerprints originated from the same source. Such conclusions, however, are dependent upon the observations of the minutia in the fingerprints.

The distinction between Levels 5–8 in HEP (the 'Conclusions') and Levels 1–4 (the 'Observations') are similar, to some extent, to the distinction between descriptive and inferential statistics, as well as to the SBAR used in the medical domain (observations on the Situation and Background, then the Assessment and Recommendation conclusions, see discussion earlier). Levels 1–4, address performance of experts in observing the data, measuring biasability and reliability between and within experts.

HEP, see Figure 1, places the observation levels below those of the conclusions because they are more basic. The observations are the ground upon which experts should reach their conclusions. When the conclusions drive the observations (rather than vice versa), confirmation bias kicks in. Therefore, conceptually and practically, observations should come first. Furthermore, observations must start with the evidence from the crime scene and then, thereafter, observations from the suspect. The danger of not following this exact sequence derives from working backwards; first, starting with the suggested conclusion (from irrelevant contextual information) and then, second, working to find the suspect's characteristics in the evidence (thus, working from the suspect to the evidence, rather than from the evidence to the suspect).

Countering these biases is relatively easy: Start with the observations (with no exposure to contextual irrelevant

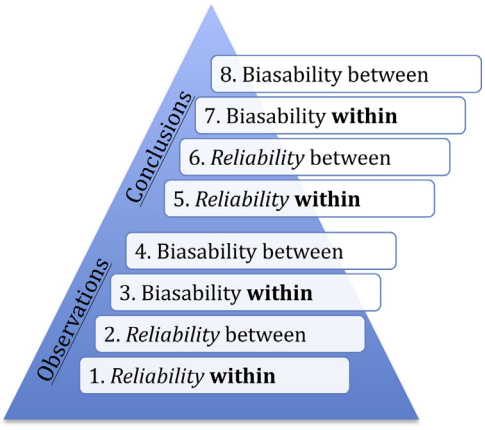


**Figure 1.** The Hierarchy of Expert Performance (HEP).

information) and work from the evidence to the suspect. The Linear Sequential Unmasking (LSU) procedure does exactly that (Dror et al., 2015). The LSU procedure mandates that forensic experts first observe the evidence from the crime scene in isolation from the suspect. They first need to examine and document what they observe in the actual evidence (e.g., in DNA, determine what are the peaks in the DNA profile from the crime scene) and, only then, be exposed to and observe the target suspect (in the DNA example, observe the peaks in the DNA profile of the suspect). Then, after observing the two DNA profiles in the correct sequence (crime scene first, then the suspect), compare them and reach conclusions (Dror et al., 2015). At no stage should the experts be exposed to irrelevant contextual information, however, as LSU permits (but with restrictions) changes can be made after exposure to the target suspect as this allows to fix inadequate initial analysis (Ulery, Hicklin, Roberts, & Buscaglia, 2014) and enables a more effective directed visual search.

Level 4, the first level in HEP quantifying observations, examines whether the observations of the data are biased by irrelevant information. One source of biasing information is the case information itself (examples of which were presented in Levels 7 and 8, such as the existence of other lines of evidence against the suspect). However, there are five different sources of biasing information (Dror, 2015; Dror et al., 2015). As discussed earlier, another source of biasing information is a 'target', the reference material to which evidence is matched (e.g., the fingerprint or DNA profile of the suspect; Thompson, 2009). Here,

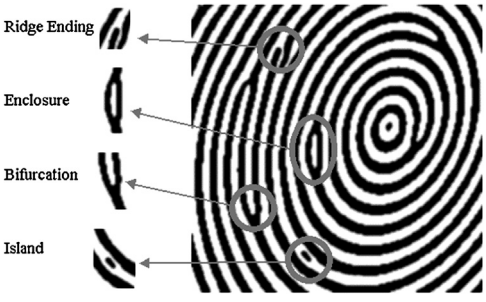


**Figure 2.** Different characteristics (minutia) present in the friction ridge of fingerprints (e.g., when a ridge divides or ends).

the evidence from the crime scene is observed within the context of the 'target suspect', which may bias how the actual evidence is observed. A study by Dror et al. (2011) examined whether such an effect exists. Forensic fingerprint experts were required to count all the minutia (see Figure 2) in a latent print from the crime scene either by itself, in isolation, or within the context of the 'target' suspect fingerprint. The results showed a biasing effect wherein the experts observed a different number of minutia when the latent print from the crime scene was presented within the 'suspect target' context.

The implications of biased observation of different numbers of minutia is not limited to subsequent comparison conclusions regarding the suspect's fingerprints, but can also affect the determination of whether the evidence is sufficient for comparison in the first place. That is, whether the latent print from the crime scene (often containing low quantity and quality of information) is even suitable for comparison. Sometimes there is insufficient data in the latent print to enable comparison with a suspect. A study by Fraser-Mackenzie, Dror, and Wertheim (2013) examined whether bias can affect the suitability determinations of latent fingerprints. In this study, experts observed latent fingerprints either in isolation or when they were presented alongside a target comparison suspect print. Similar to the study in which experts were biased and observed different number of minutia, this study demonstrated bias in the observation of whether the latent prints were suitable for comparison.

Whereas Level 4 examines the biasability in observations between different experts, Level 3 in HEP examines a more basic measure of biasability, the *within*-observations biasability of experts. That is, will the *same* expert observe the evidence differently if it is presented within irrelevant contextual information?

A research study by Earwaker, Morgan, Harris, and Hall (2015) found evidence for this type of within-expert biased observation. Observation of whether the same latent fingerprints had sufficient information for comparison was biased by irrelevant contextual information. Latent fingerprints that were insufficient and not suitable for comparison were observed as more sufficient when they were presented within the context of a serious crime (i.e., murder), whereas latent fingerprints that were sufficient and suitable for comparison were observed as insufficient when presented within the context of a less serious crime (i.e., car theft). Because of this biasing information, 34% of the latent fingerprints that were observed as sufficient were actually of insufficient quality to be compared and 33% of the latent fingerprints that were observed as insufficient and not suitable for comparison were actually of good enough quality to have been compared (Earwaker et al., 2015).

The danger of biased observation of the latent prints is illustrated in the criminal case R. v. Smith (2011) where the Court of Appeal of England and Wales quashed a murder conviction because fingerprint evidence was biased as a result of a "target" murder suspect affecting the observation of the latent fingerprint evidence. When the latent fingerprint evidence from the crime scene was examined in isolation (prior to having a target suspect), the fingerprint expert observed that "there was insufficient detail to be able to make a meaningful comparison," but then

**Table 1**
*The Between-Experts Observations of Fingerprints*

| Expert | Latent fingerprint | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | J |
| Expert 1 | 22 | 9 | 15 | 8 | 9 | 3 | 8 | 11 | 7 | 10 |
| Expert 2 | 21 | 11 | 25 | 7 | 10 | 9 | 9 | 10 | 6 | 5 |
| Expert 3 | 19 | 9 | 18 | 10 | 7 | 9 | 15 | 19 | 6 | 6 |
| Expert 4 | 21 | 21 | 29 | 14 | 12 | 9 | 8 | 9 | 4 | 8 |
| Expert 5 | 17 | 16 | 15 | 11 | 16 | 9 | 7 | 12 | 5 | 5 |
| Expert 6 | 20 | 14 | 22 | 9 | 10 | 7 | 13 | 18 | 7 | 9 |
| Expert 7 | 22 | 17 | 15 | 10 | 10 | 8 | 11 | 24 | 8 | 11 |
| Expert 8 | 9 | 9 | 19 | 6 | 9 | 8 | 18 | 16 | 9 | 10 |
| Expert 9 | 30 | 15 | 25 | 10 | 12 | 12 | 19 | 22 | 12 | 17 |
| Expert 10 | 25 | 13 | 18 | 13 | 12 | 10 | 13 | 15 | 7 | 10 |
| Min | 9 | 9 | 15 | 6 | 7 | 3 | 7 | 9 | 4 | 5 |
| Max | 30 | 21 | 29 | 14 | 16 | 12 | 19 | 24 | 12 | 17 |
| SD | 5.49 | 4.01 | 4.93 | 2.49 | 2.45 | 2.32 | 4.25 | 5.15 | 2.23 | 3.54 |
| Range | 21 | 12 | 14 | 8 | 9 | 9 | 12 | 15 | 8 | 12 |

*Note*: Different fingerprint experts looking at the same fingerprints (A to J) lack consistency in the number of minutia they observe.

"after the appellant had been charged. . he concluded that the ridge flow and 12 ridge characteristics could be identified with the fingerprint from the appellant's left forefinger." That is, the examiner "revised" his observation of the latent print, changing from insufficient and unsuitable for comparison to suitable (and then proceed to identify the suspect) (R. v. Smith, 2011, paragraphs 14 and 15). The danger derives from working backwards, from the suspect to the evidence, rather than from the evidence to the suspect (this can be addressed by the Linear Sequential Unmasking (LSU) procedure, discussed earlier).

As we move down HEP, we get to the very basic measurement of expert performance. Level 2 pertains to the basic reliability between experts (without contextual bias) in the observation of the data. Would different experts, examining identical fingerprints, observe the same number of minutia when presented without biasing context (see Levels 3 and 4 for presentation within a context)? Dror et al. (2011) investigated this question. Table 1 presents the lack of reliability and consistency among fingerprint experts.

As the data in Table 1 shows, different experts observe different data from identical fingerprints, even when presented with no biasing contextual information. This is critical, not only because of the high variability (wide range and high Standard Deviations in the number of minutia they observe), but also because the observations vary across the critical thresholds for making an identification. Although many countries do not have a set number of minutia required for identification, 7 or fewer minutia would very rarely (if ever) be sufficient for making an identification. Therefore these data (see Table 1) are alarming, as fingerprints D, E, F, G, I and J (more than half the prints) include experts who observe 7 or fewer minutia, while other experts, observing the same fingerprints, observe significantly more minutia (at least 12 minutia). Thus, the potential to make an identification conclusion is determined by which expert observes the latent print from the crime scene – for example, if Expert 2 looked at fingerprint J, then an identification would not be possible with only five minutia, however if Expert 9 looked at the very same fingerprint, then an identification would be very likely with 17 minutia. Variations between expert observations of minutia have also been quantified and characterized by Swofford, Steffan, Warner, Bridge and Salyards (2013).

Level 1, the very bottom of HEP, examines the most basic of expert performance: Whether the same expert, looking at the same evidence, will observe the same data. Such a *within*-expert measure touches upon the very basis of expertise. Dror et al. (2011) investigated this question and found that not only are forensic fingerprint experts not consistent with the observations of one another (Level 2), but the same forensic fingerprint expert looking at the same print is not always consistent with their own observations (see Table 2).

That is, the same expert observed different data from an identical fingerprint, even when presented with no biasing contextual information. These data show that the difference in the number minutia observed by the same expert at Time 1 and at Time 2 go up to nine. Reliability of within-expert performance (consistency of zero difference between Time 1 and Time 2) was observed only 16% of the time. The significance of this lack of reliability is especially alarming when it crosses the threshold for making an identification, which the data showed across experts and across the fingerprints (e.g., Expert 2 Fingerprint H, 8 vs. 13 (at one time observed 8 minutia, but at another time observed 13 minutia); Expert 3 Fingerprint G, 8 vs. 17; Expert 4 Fingerprint D, 6 vs. 11; Expert 6 Fingerprint F, 7 vs. 13; and Expert 7 Fingerprint B, 9 vs. 13 – see Dror et al., 2011). Hence, the possibility for an identification not only depends upon which expert observes the latent fingerprint (Level 2 in HEP) but, even when the same expert is involved, at one time they might observe sufficient number of minutia to enable an identification, whereas, at another time a different, much lower number, that does not enable an identification. Variations within expert observations of minutia have also been quantified and characterized by Swofford, Steffan, Warner, Bridge and Salyards (2013).

In conclusion, the Hierarchy of Expert Performance (HEP) provides a framework to examine and quantify expert

**Table 2**
*The Within-Expert Observation of Fingerprints*

| Expert | Latent fingerprint | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | J |
| Expert 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 2 | 0 | 1 |
| Expert 2 | 8 | 3 | 5 | 1 | 1 | 2 | 2 | 5 | 2 | 2 |
| Expert 3 | 1 | 3 | 3 | 3 | 6 | 4 | 9 | 9 | 1 | 2 |
| Expert 4 | 2 | 3 | 2 | 5 | 0 | 1 | 1 | 0 | 0 | 1 |
| Expert 5 | 6 | 2 | 2 | 3 | 4 | 1 | 3 | 3 | 0 | 3 |
| Expert 6 | 9 | 4 | 2 | 1 | 4 | 6 | 0 | 5 | 1 | 1 |
| Expert 7 | 0 | 4 | 5 | 2 | 4 | 3 | 3 | 7 | 0 | 0 |
| Expert 8 | 3 | 1 | 4 | 0 | 6 | 2 | 1 | 4 | 2 | 0 |
| Expert 9 | 4 | 3 | 9 | 0 | 4 | 4 | 3 | 1 | 1 | 3 |
| Expert 10 | 1 | 0 | 0 | 1 | 4 | 1 | 4 | 1 | 0 | 0 |
| MEAN | 3.5 | 2.4 | 3.6 | 1.7 | 3.4 | 2.6 | 2.9 | 3.7 | 0.7 | 1.3 |

*Note*: The difference between the number of minutia that each expert observes when examining the same fingerprint (A to J) on two separate occasions (i.e. 0 indicates consistency).

performance. First, with respect to observing the data and second, with respect to drawing conclusions. The hierarchy quantifies the extent that experts are reliable at the basic level of being consistent as well as at higher level of being biasable by irrelevant contextual information. The hierarchy also distinguishes when performance varies *between* experts and at a more basic measure of when performance varies *within* experts. This framework clearly defines eight levels, each of which quantifies a different aspect of expert performance.

The HEP framework is used to unpack and measure forensic expert performance. Since the criticisms of forensic expertise by the National Academy of Sciences (2009), there has been a growing amount of research on forensic expert performance (Found, 2015). Using HEP, these research studies can be organized and conceptualized, giving a clear theoretical framework, as well as applied implications to identify weaknesses and procedures to address them.

Although the hierarchy has been used in this paper within the forensic expert domain, it is also applicable to other expert domains. HEP may also be used to evaluate expertise by comparing performance at the 8-levels of the hierarchy across expert domains, as well as to novices. The hierarchy focuses on reliability and biasability issues that cut across expert domains, and does not address the issue of validity, which is domain dependent. Finally, this hierarchy provides practical, as well as theoretical, insights into expertise.

## Conflict of Interest Statement

The author declares no conflict of interest.

## References


Butt, L. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions—Commentary by a forensic examiner. *Journal of Applied Research in Memory and Cognition*, *2*(1), 44–45.

Coble, M. (2015). *Interpretation errors detected in a NIST inter-laboratory study on DNA mixture interpretation in the U.S (MIX13).* Washington, DC: Presentation at the international symposium on forensic science error management – detection, measurement and mitigation.

Dror, I. E. (2015). Cognitive neuroscience in forensic science: Understanding and utilizing the human element. *Philosophical Transaction of the Royal Society B*, *370*, 20140255.

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*, 10–17.

Dror, I. E., & Charlton, D. (2006). Why experts make errors. *Journal of Forensic Identification*, *56*(4), 600–616.

Dror, I. E., Charlton, D., & Peron, A. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, *156*(1), 74–78.

Dror, I. E., & Cole, S. (2010). The vision in 'blind' justice: Expert perception, judgment and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review*, *17*(2), 161–167.

Dror, I. E., & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. *Science & Justice*, *51*(4), 204–208.

Dror, I. E., & Rosenthal, R. (2008). Meta-analytically quantifying the reliability and biasability of forensic experts. *Journal of Forensic Sciences*, *53*(4), 900–903.

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*(4), 1111–1112.

Earwaker, H., Morgan, R. M., Harris, A. J., & Hall, L. A. (2015). Fingermark submission decision-making within a UK fingerprint laboratory: Do experts get the marks that they need? *Science & Justice*, *55*(4), 239–247.

Feldon, D. F. (2007). The implications of research on expertise for curriculum and pedagogy. *Educational Psychology Review*, *19*(2), 91–110.

Forensic Regulator. (2015). *Cognitive bias effects relevant to forensic science examinations*.. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/470549/FSR-G-217_Cognitive_bias_appendix.pdf

Found, B. (2015). Deciphering the human condition: the rise of cognitive forensics. *Australian Journal of Forensic Sciences*, *47*(4), 386–401.

Fraser-Mackenzie, P., Dror, I. E., & Wertheim, K. (2013). Cognitive and contextual influences in determination of latent fingerprint suitability for identification judgments. *Science & Justice*, *53*(2), 144–153.

Hoffman, R. R. (1996). How can expertise be defined? Implications of research from cognitive psychology. In R. Williams, W. Faulkner, & J. Fleck (Eds.), *Exploring expertise* (pp. 81–100). Edinburgh, Scotland: University of Edinburgh Press.

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.

Majamaa, H., & Ytti, A. (1996). A survey of the conclusions drawn of similar footwear cases in various crime laboratories. *Forensic Science International*, *82*, 109–120.

Murrie, D., Boccaccini, M., Guarnera, L., & Rufino, K. (2013). Are forensic experts biased by the side that retained them? *Psychological Science*, *24*, 1889–1897.

National Academy of Sciences. (2009). *Strengthening forensic science in the United States: A path forward*. Washington, DC: National Academies Press.

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

Smith, R. v. (2011). *EWCA (Crim) 1296 (Eng.). Case No: 2009/03393/C1. In The High Court Of Justice, Court Of Appeal of England and Wales (Criminal Division)*.

Starr, D. (2016). When DNA is lying. *Science*, *351*, 1133–1136.

Servick, K. (2015). Forensic labs explore blind testing to prevent errors. *Science*, *349*(6247), 462–463.

Spinney, L. (2010). Science in court: The fine print. *Nature*, *464*, 344–346.

Swofford, H., Steffan, S., Warner, G., Bridge, C., & Salyards, J. (2013). Inter- and intra-examiner variation in the detection of friction ridge skin minutiae. *Journal of Forensic Identification*, *63*(5), 553–570.

Thomas, C. M., Bertram, E., & Johnson, D. (2009). The SBAR communication technique. *Nurse Educator*, *34*(4), 176–180.

Thompson, W. C. (2009). Painting the target around the matching profile: The Texas sharpshooter fallacy in forensic DNA interpretation. *Law, Probability and Risk*, *8*, 257–276.

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.

Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. (2014). Changes in latent fingerprint examiners' markup between analysis and comparison. *Forensic Science International*, *247*, 54–61.

Wacogne, I., & Diwakar, V. (2010). Handover and note-keeping: the SBAR approach. *Clinical Risk*, *16*(5), 173–175.

Received 15 March 2016;
accepted 16 March 2016
Available online 24 March 2016