J Forensic Sci, July 2008, Vol. 53, No. 4
doi: 10.1111/j.1556-4029.2008.00762.x
Available online at: www.blackwell-synergy.com

Itiel Dror,[1] *Ph.D. and Robert Rosenthal,*[2] *Ph.D.*

Meta-analytically Quantifying the Reliability and Biasability of Forensic Experts

ABSTRACT: In this paper we employ meta-analytic procedures and estimate effect sizes indexing the degree of reliability and biasability of forensic experts. The data are based on within-expert comparisons, whereby the same expert unknowingly makes judgments on the same data at different times. This allows us to take robust measurements and conduct analyses that compare variances within the same experts, and thus to carefully quantify the degree of consistency and objectivity that underlie expert performance and decision making. To achieve consistency, experts must be reliable, at least in the very basic sense that an expert makes the same decision when the same data are presented in the same circumstances, and thus be consistent with themselves. To achieve objectivity, experts must focus only on the data and ignore irrelevant information, and thus be unbiasable by extraneous context. The analyses show that experts are not totally reliable nor are they unbiasable. These findings are based on fingerprint experts decision making, but because this domain is so well established, they apply equally well (if not more) to all other less established forensic domains.

KEYWORDS: forensic science, fingerprint identification, confirmation bias, psychological influence, cognitive top-down processing, forensic decision making, reliability and biasability of expert decision making

Human judgments are affected by a variety of factors. These effects stem from our initial perceptual mechanisms to higher cognitive functions (1). Given such variability and individual differences, the question arises: How reliable are human judgments? The variability across individuals reflects that people are different; they have different past experiences, mental representations and abilities, as well as different motivations, personalities, and so forth.

This issue becomes especially acute when considering judgments made by experts. Whereas the possible lack of reliability of judgments in everyday life may be acceptable (and even warranted), the lack of reliability in expert domains, such as medicine and the criminal justice system, is especially concerning. For example, the notion that whether a suspect is found guilty or innocent actually depends on the specific judge they face, is problematic and unacceptable in principle. We require that expert judges rule on the actual innocence or guilt of the suspect, or at the very least on the evidence for their guilt. The possibility that some judges would find a suspect innocent whereas other judges would find the same suspect guilty stands in sharp contradiction to the basic foundations of the criminal justice system.

Although between-expert reliability can vary, what about within-expert reliability? In such cases, different judgments would not be made across different experts, but different judgments would be made by the very same expert on the very same situation. This addresses the most basic and pure notion of reliability: Would an expert make the same judgment when presented with the same case? It would be more than troubling if the same medical expert would sometimes recommend a serious and dangerous operation and at other times would recommend against it, when judging an identical situation.

A lack of within-expert reliability is much more of a concern than a lack of between expert reliability. The latter may reflect individual differences, and may even be, depending on the domain, an advantage at providing a multitude of viewpoints and analyses. However, the former represents basic inconsistency and a failure in judgments. If experts are not reliable in the sense that they are not consistent with themselves, then the basis of their judgments and professionalism is in question. From a scientific and research point of view we are not interested or concerned with cases in which experts are not reliable because of lack of attention, negligence, or dishonesty [see (2), for classification of different expert errors]. It is important to study the performance of dedicated and competent experts and examine their reliability in the real world at the very basic level of being consistent with themselves.

Another important perspective in expert decision making is whether they are biasable in their judgments. Experts, as humans, perceive and judge information based on circumstances, such as context, emotional states, expectations, and hopes. Circumstances constantly change, as Heraclitus allegedly said, "you could not step twice in the same river" (in fact this is actually what Plato said, reconstructing—and distorting—Heraclitus's position: "Heraclitus, I believe, says that all things go and nothing stays, and comparing existents [*sic*] to the flow of a river, he says you could not step twice into the same river" *Cratylus* 402a = DK22A6, [3]). Thus, experts faced with different circumstances may reach different and even conflicting judgments and conclusions.

This is not a problem if the circumstances are relevant to their decision making, because by being relevant the new circumstances may actually change the decision problem itself. But what happens when experts are faced with extraneous circumstances which are not relevant and do not modify the decision problem? Would those bias and contaminate their professional and "objective" judgments? It is important to consider not only whether experts are reliable, but also what factors may affect and bias their professional judgments.

Although it is critical to empirically and properly study the reliability and biasability of experts, this type of research is extremely scarce and almost nonexistent. One reason for this is the difficulty of conducting empirical research in this area; it is extremely

[1]School of Psychology, University of Southampton, Southampton, SO17 1BJ, United Kingdom.

[2]Department of Psychology, University of California, Riverside, CA 92521.

Received 9 Aug. 2007; and in revised form 18 Nov. 2007; accepted 21 Nov. 2007.

challenging on three fronts—accessibility, ecological validity, and data analysis—as we specify below.

In general, experts' time and availability is a very limited and expensive resource. Nevertheless, it is relatively straightforward to bring experts into a research laboratory to assess, compare, and characterize their performance and abilities (e.g., [4,5]). However, experts and their organizations are apprehensive about being studied and examined. This understandably is drastically increased when the study is examining sources of expert error, reliability, and bias. Such studies question expert performance and, by their very nature, elicit resistance and defensiveness, and therefore make it especially challenging to get consent or volunteers to participate. On top of that, there are legal issues involved in demonstrating expert error in real domains and cases such as medicine and the criminal justice system.

Even if accessibility is resolved, the experimental procedures need to assure that participants consent to participate in the study, but yet are unaware when the data are collected. Experts' performance in laboratory conditions does not reflect their performance in their day-to-day activities in real life circumstances. Even in field studies, experts behave differently when they are observed, and especially if the study is examining their reliability and biasability. Thus, one needs to access the experts in their normal environment and everyday casework, but without their knowledge, and yet conduct experimental manipulations and data collection as similarly as possible to laboratory conditions. Furthermore, the framework of scientific research requires an open mind and personal motivation to find the truth; however people involved in the criminal justice system work and many times think within the adversarial legal system whereby people have a motivation to prove an *a priori* position. These issues make it hard to apply scientific methodology and research to study errors within forensic experts.

In addition, one of the greatest challenges in conducting studies that examine within-expert reliability and biasability is that we need to use the experts as their own control and thus use a repeated measures design. We must have the experts judge the same case twice, so we can compare their consistency across the two presentations. But any awareness that they have seen this case in the past may invalidate the results.

Given the great obstacles to conducting these studies, it is extremely difficult to carry out this type of research. When the above stipulations are followed, they pose such difficulties that the data are likely to be very scarce. With small samples effect sizes can only be imprecisely estimated, and tests of significance are made with low statistical power. As far as we know, there are only two studies that have examined expert reliability and biasability performance and conformed to all the stipulations we listed above. However, these studies come up short in dealing with the limits of the data analysis. Accordingly they can only provide clues to the existence and the magnitude of the problem. In this paper, we want to take steps to remedy the above limitations and to make a contribution to methods of studying reliability and biasability of experts at the basic level of within-expert performance. Thus, we apply meta-analytic procedures, and employ a statistical tool that enables us more accurately to estimate effect sizes in small samples, the $r_{\text{equivalent}}$ (6).

The statistic, $r_{\text{equivalent}}$, is an effect size estimate that can be computed knowing only a *p*-value and a sample size. There are three major situations for which $r_{\text{equivalent}}$ was especially designed: (a) in meta-analytic work, or in other re-analyses of other people's data, when neither effect sizes nor test statistics (e.g., *t*, *F*, χ^2) are provided; (b) no effect size estimate exists for the data-analytic procedures used (e.g., sign test; Wilcoxon–Mann–Whitney test, Kolmogorov–Smirnov test, permutation tests); and (c) an effect size estimate *could* be computed directly from the data but, because of small sample sizes (or severe non-normality), the directly computed estimates may be seriously misleading.

In the present study, we have a small sample size (as explained above, this is inherent to these types of studies). If we directly computed r_{sample} based on $r_{\text{sample}} = \sqrt{\chi^2_{(1)}/N}$ then this can be seriously inflated and misguiding. Since r_{sample} is defined and computed using $\chi^2_{(1)}$, r_{sample} is too large when $\chi^2_{(1)}$ is based on small expected frequencies, with "small" often given as less than 5 per cell. None of the 2 × 2 tables of results for each expert (see Table 1) come close to having all four expected frequencies at least 5 per cell, with none of the 24 cells having expected frequencies larger than 2.50. Therefore, under this situation ("c", above) *p*-values based on Fisher's exact test are more accurate than those based

TABLE 1—*Six experts' retest reliability of judging fingerprint matches.*

		Original					
Expert	Retest	Match	No match	Fisher Exact *p*	*t* from *p**	r_{sample}	$r_{\text{equivalent}}$
A	Match	4	0	0.014	2.87	1.00	0.76
	No match	0	4				
B	Match	3	0	0.071	1.69	0.77	0.57
	No match	1	4				
C	Match	3	0	0.071	1.69	0.77	0.57
	No match	1	4				
D	Match	2	1	0.43	0.18	0.26	0.07
	No match	2	3				
E	Match	3	0	0.029	2.46	1.00	0.74
	No match	0	4				
F	Match	4	0	0.014	2.87	1.00	0.76
	No match	0	4				
	Median			0.050	2.07	0.88	0.66
Combined experts	Match	19	1	$2.2/10^8$	6.58	0.79†	0.70‡,§
	No match	4	23				

**df* = 6 for each expert's *t* except for expert E who lost 1 *df* by omitting a decision on one of his or her trials.

†From $\sqrt{\chi^2_{(1)}/N} = \frac{Z}{\sqrt{N}}$.

‡From Fisher Exact *p*.

§95% confidence interval runs (in units of *r*) from 0.52 to 0.82 for a fixed effect (*N* = 47 decisions) and from 0.33 to 0.80 for a random effect (*N* = 6 experts).

on $\chi^2_{(1)}$, and since $r_{\text{equivalent}}$ is based on p-values derived from Fisher's exact test, $r_{\text{equivalent}}$ will be more accurate than r_{sample}.

We briefly describe the first study, report the p-value obtained, and estimate the effect size for reliability. Then we describe the second study and meta-analytically combine the results of the two studies on biasability.

Reliability

Past decisions of fingerprint experts were retrieved from archives of real criminal cases and then re-presented to the *same* experts. These experts agreed to take part as participants in research, and are on our database of international experts willing to take part in our studies. To be put on the database you must be a qualified expert with accredited training and substantial experience (this varies from country to country, as training durations and qualifications vary). Furthermore, if you consent to participate in our research, you consent that in the next 5 years we may use you in our studies, without your knowledge. We further stipulate that we will not use many of the experts in our database, thus the experts do not know if and when they take part in the study.

The first study consisted of six experts. Each of these experts had more than 5 years' experience in latent fingerprint examination post training and accreditation. They all consistently passed all their proficiency testing and were regarded as very competent examiners by their supervisors. Each expert was approached by their manager/supervisor and asked to make eight judgments on pairs of prints (without being told that this is a study or that they have previously made judgments on these exact prints). The fingerprints were given to the experts in the same format that they were presented to them in the past. Four of those judgments were given to address the basic and pure notion of reliability that we have discussed: Would experts make the same judgment on the same decision problem? The other four judgments were given to address issues of biasability, which are dealt with in the next section. These data are reported purely as descriptive data elsewhere (2). Here we attempt to make sense of the data by computing p-values and examining effect sizes, using statistical tools that provide a more accurate assessment of the data even with such a small sample.

Analysis and Discussion: Reliability

Table 1 shows the retest reliability for each of the six experts, individually and combined. Table 2 shows the back-to-back stem and leaf displays of directly computed reliabilities (r_{sample}) and the small-sample-size-adjusted reliabilities ($r_{\text{equivalent}}$).

Based on $r_{\text{equivalent}}$ we see the 95% confidence interval around the mean retest reliability of six fingerprint experts ranges from 0.33 to 0.80, a good deal lower than the value we might have expected were fingerprint experts to be nearly always accurate. Even if we consider the more "optimistic" indices of reliability, r_{sample}, we find three of our six experts with retest reliabilities of 0.77 or below.

When we consider the $r_{\text{equivalent}}$ associated with the total number of 47 judgments that could be considered as clearly agreeing or clearly disagreeing with their own earlier judgments, our 95% confidence interval does narrow noticeably, ranging now from a low of 0.52 to a high of 0.82. None of the retest reliabilities in this 95% confidence interval range are high enough to reassure us of the overall accuracy or consistency of our fingerprint experts.

To provide a balanced view, however, we must emphasize that our fingerprint experts do indeed show considerable expertise in the sense of accuracy (or at least consistency) beyond chance levels. Indeed, we could say that for the 47 retest trials that could be categorized as quite right or quite wrong, 89% were quite right. That is far better than chance ($p = 2.2/10^8$) but still very far from what is the accepted and recognized norm of fingerprint expert performance.

TABLE 2—*Back-to-back stem and leaf displays of fingerprint experts' retest reliability.*

r_{sample}*				$r_{\text{equivalent}}$†
N = 6 experts				*N = 6 experts*
Median: 0.88	0, 0, 0	1.0		Median: 0.66
Mean: 0.80		0.9		Mean: 0.58
Range: 0.26–1.00		0.8		Range: 0.07–0.76
$S_{r_{\text{sample}}}$: 0.29	7, 7	0.7	4, 6, 6	$S_{r_{\text{equivalent}}}$: 0.26
		0.6		95% CI: 0.33–0.80
N = 47 judgments		0.5	7, 7	N = 47 judgments
r_{sample}: 0.79		0.4		$r_{\text{equivalent}}$: 0.70
		0.3		95% CI: 0.52–0.82
	6	0.2		
		0.1		
		0.0	7‡	

*Directly computed r from $Z/\sqrt{N}$.

†Small-sample-size-adjusted r based on Fisher exact test p-value.

‡This expert's $r_{\text{equivalent}}$ of 0.07 is not quite an outlier ($M_{\text{i}} = 2.8$) using Iglewicz and Hoaglin's criterion of $M_{\text{i}} > 3.5$ (7). In addition, the six values of $r_{\text{equivalent}}$ showed no unusual heterogeneity, $\chi^2_{(5)} = 3.16$, $p = 0.68$, $S = 0.36$ based on Fisher's Z_r transformation ([8] p. 74).

Biasability

The first study also included four judgments from each of the six experts in which context was manipulated so as to examine whether extraneous information might bias the expert fingerprint examiners in their judgments leading them to make different and conflicting decisions to those they made in the past on the very same fingerprints. This extraneous context included information such as "the suspect has an alibi" (for manipulations aimed at biasing the expert to judge the prints as a nonmatch), "the suspect confessed to the crime" (for manipulations aimed at biasing the experts to find a match).

In another study five other experts were tested on a single decision problem. The experts in this study were also taken from our database of international fingerprint experts willing to take part in our studies, and followed the criteria from the other study, specified above. In this other study, each of the experts needed to decide whether a pair of fingerprints matched, and they were provided with a context that could bias them to decide the fingerprints were not a match. Not only did the fingerprints match, but again, the prints were in fact judged by the same experts as a match in the past. Due to the small data set and sample size, these data too were only reported descriptively (9).

In this section we first analyse each study separately, computing both p-values and appropriate effect size estimates. After analysis of each study separately, we combine the results of both studies meta-analytically to obtain greater statistical power and more accurate effect size estimates.

Analysis and Discussion: Biasability

For one of the studies there were six experts, each making eight judgments. The four judgments that had no contextual manipulation examined reliability *per se*, and were analysed and discussed in the Reliability section. In this section, we focus on the four judgments that included an experimental manipulation of biasing instructions (extraneous contextual information). The data showed a substantial magnitude of effect ($r = 0.41$, $p = 0.18$) with three of the six

TABLE 3—*Binomial effect size display of an effect size r of 0.40.*

	Judgment Accuracy		
	Correct	Incorrect	Σ
Contextual information			
Bias	30	70	100
No bias	70	30	100
Σ	100	100	200

experts affected by the biasing instructions. For the small N involved, however, the magnitude of the effect showed a very wide 95% confidence interval (from $r = -0.60$ to $r = 0.92$). However, lest we conclude that the actual magnitude of the biasing effect was zero, we note that the counternull value of the obtained effect size r was 0.70 which means the population value of the effect size r has exactly the same probability of being 0.70 as it does of being 0.00, i.e., the null value of r (10).

In the other study there was a single sample of five experts, all of whom were given biased contextual information that affected four of the five experts, i.e., led them to change from an earlier judgment of match (that had been corroborated by additional experts) to a later (probably incorrect) judgment of no match, or "can't tell." A very conservative binomial test, hypothesizing as a null only that biased instructions would lead equally often to match and to no match judgments, yielded $r_{\text{equivalent}} = 0.45$, very close to the value of $r = 0.41$ obtained in the more recent study, but like that study, showing a very wide 95% confidence interval (from $r = -0.72$ to $r = 0.95$). With confidence intervals so wide, we might be inclined to believe that the actual magnitude of the biasing effect was zero; however, the counternull value of our effect size r was 0.74 which means the population value of the effect size r has exactly the same probability of being 0.74 as it does of being 0.00.

Taking the effect size r of both studies (rs = 0.41, 0.45) we find a one sample $t_{(1)}$ of 21.5, $p = 0.015$. The magnitude of the biasing effect of extraneous contextual information reminds us that biasing of fingerprint experts is not a minor problem. Viewing the approximate size of the biasing effect by means of a binomial effect size display (BESD) reveals that the practical importance of an r of 0.40 can be seen in Table 3 as increasing the proportion of judgments that are incorrect from 30% to 70% (11).

General Discussion

Collecting data covertly in the field is challenging but necessary if we want to observe and understand how experts really go about their business. To examine reliability and biasability in the purest fashion, the data consisted of repeated measures whereby the experts and the stimuli were their own control. Thus, experts were tested on cases they had judged in the past. To examine reliability, the stimuli were re-presented to the experts to examine and judge. We wanted to learn the degree to which they would make the same or conflicting decisions compared to their past judgments. To examine biasability, the context was manipulated when the stimuli were re-presented. Extraneous information was added, so as to cause the experts to have expectations about the outcome of the examination. Here it was of interest whether the experts would be able to make the judgment based on the stimuli alone, or whether they would be biased in their perception and interpretation because of the extraneous context provided.

Because of the nature of these studies, they included small sample sizes (five and six experts) and limited data sets: eight outcomes (fixed effect) or one outcome (random effect) per expert. Thus, the studies had been reported only descriptively. In this paper we examined the data further. First we computed p-values and effect size estimates. Because of the very small sample sizes, effect size estimates can be exaggerated and quite misleading. Hence we employed an effect size estimate, $r_{\text{equivalent}}$, which takes into account the sample size and thus provides a more appropriate effect size estimate. We also combined the data from both studies meta-analytically to increase statistical power and improve the accuracy of our estimates of effect sizes.

Are fingerprint experts reliable? Are they unbiasable? The first two studies to examine these questions established that experts are far from being perfect. These studies demonstrated circumstances in which experts were both relatively unreliable and biasable, and in the analyses reported here we quantify these effects statistically and subject them to meta-analytic procedures. The data are based on forensic decision making made by latent fingerprint experts, but because this forensic domain is the most widely used and well established, we can be confident that the problems exposed within this domain are also prevalent in other forensic domains.

The fact that fingerprint experts can be unreliable and biasable does not mean that they are not ordinarily reliable and unbiasable. It is not our place to determine what is the acceptable norm for expert performance, in this or any other forensic domain. We do however develop and provide the experimental methodology and quantitative statistical tools to examine and quantify their performance, specifically in terms of reliability and biasability. Such quantification is critical for determining acceptable norms of expert performance, as well as developing and evaluating training and procedures to improve such performance. Of course, using scientific studies for advancing the domain depends on the openness of the forensic community to take such findings on board and not engage in defensive responses and denial.

References

1. Dror IE. Perception is far from perfection: the role of the brain and mind in constructing realities. Brain Behav Sci 2005;28(6):763.
2. Dror IE, Charlton D. Why experts make errors. J Forensic Ident 2006;56(4):600–16.
3. Diels H, Kranz W. Die Fragmente der Vorsokratiker. Zurich: Weidmann, 1985.
4. Dror IE, Kosslyn SM, Waag W. Visual-spatial abilities of pilots. J Appl Psychol 1993;78(5):763–73.
5. Dror IE. The effects of screening, training, and experience of air force fighter pilots: the plasticity of the ability to extrapolate and track multiple objects in motion. North Am J Psychol 2004;6(2):239–52.
6. Rosenthal R, Rubin DB. $r_{\text{equivalent}}$: a simple effect size indicator. Psychol Methods 2003;8:492–6.
7. Iglewicz B, Hoaglin DC. How to detect and handle outliers. Milwaukee: ASQC Quality Press, 1993.
8. Rosenthal R. Meta-analytic procedures for social research. Newbury Park, CA: Sage, 1991.
9. Dror IE, Charlton D, Peron A. Contextual information renders experts vulnerable to making erroneous identifications. Forensic Sci Int 2006;156(1):74–8.
10. Rosenthal R, Rubin DB. The counternull value of an effect size: a new statistic. Psychol Sci 1994;5:329–34.
11. Rosenthal R, Rubin DB. A simple general purpose display of magnitude of experimental effect. J Educ Psychol 1982;74:166–9.

Additional information and reprint requests:
Itiel Dror, Ph.D.
School of Psychology
University of Southampton
Southampton SO17 1BJ
United Kingdom
E-mail: id@ecs.soton.ac.uk
URL: http://users.ecs.soton.ac.uk/id/biometrics.html