2. Unreasonable expectations of representative data

It would certainly be desirable to conduct studies based on representative samples of fingerprint data, representative latent print examiners, and standard operating procedures shared among all laboratories in the country. Unfortunately, this is not a realistic requirement for studies that involve multiple agencies. In BB we stated

> *There is substantial variability in the attributes of latent prints, in the capabilities of latent print examiners, in the types of casework received by agencies, and the procedures used among agencies. Average measures of performance across this heterogeneous population are of limited value* [...]*—but do provide insight necessary to understand the problem and scope future work. Furthermore, there are currently no means by which all latent print examiners in the United States could be enumerated or used as the basis for sampling: A representative sample of latent print examiners or casework is impracticable*[...].

Haber and Haber state "the experimental designs employed deviated from casework procedures in critical ways." [Haber and Haber, Abstract] While a representative sampling of casework from a given agency might be practical, given the variability of standard operating procedures among agencies (detailed in [BB SI-1.4]), it would not be possible for ANY study to be representative of casework across a variety of agencies. BB results provide an example of how it is unrealistic to assume that a study could be based on fully representative data selection: four of the five latents that resulted in false positives were on galvanized metal, processed with cyanoacrylate and light gray powder. We received multiple comments from examiners who said that they had never seen such prints. In casework, the frequency of prints with such characteristics may vary significantly among agencies.

3. Accounting for inconclusive decisions

Haber and Haber consider all inconclusives as "missed identifications." "Missed identification" can be a misleading term and is not used consistently by the latent print community, but typically the term connotes a failure on the part of a single examiner to individualize when other examiners deem that individualization is justified. Haber and Haber discuss missed identifications several times, defining all instances of non-identified mated data as missed identifications. Haber and Haber are including cases where examiners unanimously agree that there is insufficient basis for individualizing, even if making an individualization in such cases could be considered reckless.

Because the Haber and Haber paper contains so many errors, we respectfully request that this letter be published in order that the inaccurate data and unsubstantiated conclusions may be corrected in the public record. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government. This is publication number 14-05 of the FBI Laboratory Division.

References

[1] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. U. S. A. 108 (19) (2011) 7733–7738 (http://www.pnas.org/content/108/19/7733.full.pdf).

[2] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Repeatability and reproducibility of decisions by latent fingerprint examiners, PLoS ONE 7 (2012) 3 http://dx.doi.org/10.1371/journal.pone.0032800.

[3] G. Langenburg, C. Champod, T. Genessay, Informing the judgments of fingerprint analysts using quality metric and statistical assessment tools, Forensic Sci. Int. 219 (1) (2012) 183–198http://dx.doi.org/10.1016/j.forsciint.2011.12.017.

[4] Z.W. Evett, R.L. Williams, Review of the 16 point fingerprint standard in England and Wales, J. Forensic Identif. 46 (1) (1996) 49–73.

[5] G. Langenburg, A performance study of the ACE-V process: a pilot study to measure the accuracy, precision, reproducibility, repeatability and bias ability of conclusions resulting from the ACE-V process, J. Forensic Identif. 59 (2) (2009) 219–257.

[6] Federal Bureau of Investigation Laboratory (1999) Survey of Law Enforcement Operations in Support of a Daubert Hearing. U.S. v. Mitchell, 365 F.3d 215 (3rd Cir.) (never published).

[7] I.E. Dror, D. Charlton, A. Peron, Contextual information renders experts vulnerable to making erroneous identifications, Forensic Sci. Int. 156 (1) (2006) 74–78.

[8] I.E. Dror, D. Charlton, Why experts make mistakes, J. Forensic Identif. 56 (4) (2006) 600–616.

[9] L.J. Hall, L. Player, Will the introduction of an emotional context affect fingerprint analysis and decision-making? Forensic Sci. Int. 181 (1) (2008) 36–39.

[10] G. Langenburg, C. Champod, P. Wertheim, Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons, J. Forensic Sci. 54 (3) (2009) 571–582.

[11] K. Wertheim, G. Langenburg, A.A. Moenssens, A report of latent print examiner accuracy during training exercises, J. Forensic Identif. 56 (1) (2006) 55–93.

[12] S. Gutowski, Error Rates in Fingerprint Examinations: The View in 2006, The Forensic Bulletin (Autumn 2006) (2006) 18–19.

[13] J.M. Tangen, M.B. Thompson, D.J. McCarthy, Identifying fingerprint expertise, Psychol. Sci. 22 (8) (2011) 995–997.

R. Austin Hicklin
Bradford T. Ulery
Noblis, United States

JoAnn Buscaglia*
Maria Antonia Roberts
FBI Laboratory, United States

*Corresponding author at: Counterterrorism and Forensic Science Research Unit, Federal Bureau of Investigation Laboratory, 2501 Investigation Parkway, Quantico, Virginia 22135, USA. Tel.: +1 703 632 4553; fax: +1 703 6324557.
E-mail address: joann.buscaglia@ic.fbi.gov (J. Buscaglia).

http://dx.doi.org/10.1016/j.scijus.2014.06.007

Generalization in fingerprint matching experiments

In their paper, "Experimental results of fingerprint comparison validity and reliability: A review and critical analysis," Haber and Haber [1] set out to determine whether 13 published studies on the performance of human fingerprint examiners can be generalized to fingerprint laboratory casework. The authors review each of the studies and clearly detail the measures they used to evaluate them. The article is, by far, the most thorough and detailed review of fingerprint identification experiments to date, and is an important contribution by this measure alone.

Haber and Haber evaluate our first published experiment on fingerprint expertise [2]. The experiment was designed to determine whether fingerprint experts are any more accurate at matching prints than lay people, and to get an idea of how often these two groups make "misses" (i.e., errors comparable with allowing a guilty person to escape detection) compared with how often they make "false alarms" (i.e., errors comparable with falsely incriminating an innocent person). We found that qualified, court-practicing fingerprint experts were exceedingly accurate at matching fingerprints compared with novices, and that experts showed a conservative

response bias. We have since conducted another experiment [3] (not cited by Haber and Haber) using genuine crime scene prints, but where the ground truth is uncertain. Again, we found that qualified court-practicing fingerprint experts were exceedingly accurate compared with novices, and that experts showed a conservative response bias.

As Haber and Haber note, Tangen et al. [2] was not designed to determine the likelihood of errors in practice, nor the performance of individuals or departments, and examiners were not provided with their usual tools or independent verification. We designed this experiment (and our subsequent experiments) to determine performance differences based on expertise using ground truth and genuine crime scene latents. Inferring from these results that experts are around 99% accurate in practice or that the overall error rate of fingerprint identification is around 1%, would be unjustified. How experts would perform with their usual tools, peer verification, statistical models, different lifting agents and surface types, different response types, time and resource constraints, different types of training, experience, and qualifications, and so forth, is still unknown.

We preempted several of the concerns raised by Haber and Haber by publishing a detailed commentary on our initial experiment [4]. In this commentary (not cited by Haber and Haber), we describe our experimental approach and methodology, as well as factors that affect generalization to casework. This paper addresses many of the concerns raised by Haber and Haber, and so we will not rebut each one in this short letter. Instead, we turn our attention to the general problem of generalization that they raised.

The intuition that the best experiment should resemble 'real-life' as far as possible is understandable, but incorrect [5]. When measuring human performance, the challenge is to effectively balance fidelity, generalizability, and control to properly address the research question at hand [4]. Fidelity is the extent to which the experimental task mirrors the reference domain, generalizability refers to the applicability of the results to circumstances beyond those examined in the experiment, and control refers to the freedom one has to isolate and manipulate variables [6]. The perfect experiment with high fidelity, high control, and high generalizability is impossible, however, so these three parameters must be balanced appropriately.

We contend that several of the "design problems" listed by Haber and Haber are not design problems at all, but factors that one may choose to consider when making an inference about matching performance during casework on the basis of formal experiments. Laboratory-based experiments, like ours, are intentionally artificial, because they allow us to systematically manipulate the factors of interest (e.g., the difference between expert and novice performance or similar and non-similar distractors) while controlling for extraneous factors that don't come apart in the wild (e.g., 'inconclusive' judgments, verification, and software tools).

Haber and Haber seem to be grappling with the very same issue that we are: what is the appropriate level of analysis and on what basis can we generalize from experimental studies to performance during casework? Our ongoing program of research is about determining the factors that affect matching accuracy, to better understand the development of expert forensic identification, to inform training, and to provide an empirical basis for expert testimony in the courtroom. Little is known about the nature and development of fingerprint expertise and, therefore, the best way to turn novices into experts. And little is known about the factors that affect matching accuracy and, therefore, what experts can legitimately testify to in court [7].

When listening to the testimony of a fingerprint examiner in court, it is tempting to think that one only needs to consider the particular prints in question, or with the opinion expressed by the examiner in this specific case. The accuracy of a particular fingerprint identification cannot be known, however. Instead, we can appeal to evidence in the aggregate about performance measures and factors that might affect the strength of the evidence in the particular case. We have suggested that a fingerprint comparison can be regarded as a diagnostic test by a human (with the aid of technology, etc.) that produces an opinion [7]. In order to make a judgment about whether to 'believe' the examiner or not, or how much weight to give her opinion, we need to know something about her prior performance and the factors that might affect her judgments—this is what our experiments, and others, can help provide.

The National Academy of Sciences [8] and others [9] have called for a culture of collaboration and a commitment to evidence from empirical research to ensure the integrity of forensics as an investigative tool. A tit for tat exposition of "design flaws" – legitimate methodological and statistical flaws and limitations notwithstanding – and outright rejection of studies is unlikely to be a fruitful approach toward our presumably shared goal of continuous improvement of forensic science systems [10]. Instead, we propose that further programs of research on the factors that affect fingerprint matching accuracy and performance, would better serve to increase our confidence about the legitimacy of claims made by expert witnesses in court, if, of course, those claims are based on the best available empirical evidence.

References

[1] R.N. Haber, L. Haber, Experimental results of fingerprint comparison validity and reliability: a review and critical analysis, Sci. Justice 54 (2014) 375–389 (in this issue).
[2] J.M. Tangen, M.B. Thompson, D.J. McCarthy, Identifying fingerprint expertise, Psychol. Sci. 22 (2011) 995–997, http://dx.doi.org/10.1177/0956797611414729.
[3] M.B. Thompson, J.M. Tangen, D.J. McCarthy, Human matching performance of genuine crime scene latent fingerprints. Law and Hum Behav, http://psycnet.apa.org/psycinfo/2013-26306-001/.
[4] M.B. Thompson, J.M. Tangen, D.J. McCarthy, Expertise in fingerprint identification, J. Forensic Sci. 58 (2011) 1519–1530, http://dx.doi.org/10.1111/1556-4029.12203.
[5] D.G. Mook, In defense of external invalidity, Am. Psychol. 38 (1983) 379–387.
[6] D. Brinberg, J.E. McGrath, Validity and the research process, SAGE Publications, California, 1985.
[7] G. Edmond, M.B. Thompson, J.M. Tangen, A guide to interpreting forensic testimony: scientific approaches to fingerprint evidence, Law Probab. Risk 13 (2013) 1–25, http://dx.doi.org/10.1093/lpr/mgt011.
[8] National Academy of Sciences, Strengthening Forensic Sciences in the United States: A Path Forward, National Academies Press, Washington DC, 2009.
[9] J.L. Mnookin, et al., The need for a research culture in the forensic sciences, UCLA Law Rev. 58 (2011) 725–775.
[10] Expert Working Group on Human Factors in Latent Print Analysis, Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach, Government Printing Office, Washington DC, 2012.

Matthew B. Thompson*
Jason M. Tangen
The University of Queensland, St Lucia, QLD 4072, Australia
*Corresponding author at: School of Psychology, The University of Queensland, St Lucia, QLD, 4072, Australia.

http://dx.doi.org/10.1016/j.scijus.2014.06.008