Letter to the Editor

Re: A Report of Latent Print Examiner Accuracy During Comparison Training Exercises, *J. For. Ident.* 56 (1).

Wertheim, Langenburg, and Moenssens (WLM) (2006) offer results from training exercises as empirical evidence of the accuracy with which examiners make fingerprint comparisons. However, due to profound methodological flaws, their results provide no better evidence for examiner accuracy or error rates than the available data they criticize at length.

WLM do not describe any purpose in their article. However, since the entire article focuses on criticizing evidence of accuracy levels of fingerprint examiners reported from other research and why the present experiment provides useful accuracy results, the reader is left to assume that the purpose of the research from the beginning was to collect data on accuracy levels of fingerprint examiners in making fingerprint comparisons.

This is a misrepresentation of the facts. The original purpose of this extensive research project was to use the participants from these training groups to develop and validate a measure of latent print difficulty. It had nothing to do with accuracy of comparisons, though assessment of errors was included as one of the measures in the development of a quantification of latent print difficulty.

An experiment to develop a validated measure of latent print difficulty was proposed to us, the writers of this Letter to the Editor (HH), by the first author of the article (W). He asked us to help him design the project, and the three of us worked out the measures and response sheets in detail jointly. We (HH) also enrolled as participants in the first training course in which these data were collected.

The design to meet the original purpose (latent print difficulty) measured one main dependent variable: seven different difficulty ratings for each latent print. Four further dependent variables were also measured to quantify a multidimensional difficulty measure for latent fingerprints, including analysis and comparison time required, correctness of any identification, confidence in the identification, and self-reported training and experience of each participant.

In contrast, the WLM report treats the major dependent variable as whether an identification was correct. This switch in the purpose after the experiment was designed and the data had been collected created severe methodological flaws. To make these evident, we fully describe the experimental procedures WLM employed.

Experimental Design of WLM

The experimenters recorded the accuracy of latent to exemplar comparisons from 108 fingerprint examiners, each of whom was a participant in one of a number of training courses on latent print comparisons sponsored by the IAI from late 2003 and 2004. These courses were open to anyone who wished to enroll and paid the fees. During a course, each participant received a various number of comparison packets, each packet consisting of ten latent prints, eight sets of inked exemplars, and a response sheet.

The latents in each packet had been previously rated for difficulty by the course instructors on a subjective basis of quality and quantity of information, focal features, distortions, and anatomical source. All of the latents were deemed of value for comparison by the instructors, and every latent print matched one of the exemplars in the packet.

The instructor determined the difficulty level of each participant's initial packet, based on the participant's self-reported training and experience. Subsequent packets for each participant were selected by the instructor on the basis of the participant's performance. The comparison exercises comprised part of the course requirement: each participant had to make a minimum of 60 correct identifications to pass this part of the course. Comparisons were done as part of class time set aside for that purpose.

When a participant first received a packet, he or she was asked to look only at the latent prints, and rate each latent on a score sheet on each of seven scales: the quality of detail, the quantity of detail, presence of focal points, level of contrast, amount of lateral distortion, amount of deposition pressure, and level of background interference. After the participant rated each of the ten latents, then he or she was to find the matching

inked print. When finished, the participant's score sheet for that packet contained the seven ratings of each of the ten latents, the time spent on each latent (combining rating and comparison time), the perpetrator's name and the number of the exemplar digit matched (if identified), and the participant's confidence of the identification on a three point scale that ranged from case work confidence to rough guess. Participants were not required to complete every latent in a packet. Further, a "stooge" in the form of the instructor provided useful hints if a participant asked for help.

When the participant turned in the packet with its score sheet, the instructor checked the identifications made for correctness and gave the participant a new packet. There was no penalty for a failure to identify any latent, or for erroneous identifications made with less than complete confidence. Only erroneous identifications made at the highest level of confidence were penalized. For every erroneous identification made at a case work level of confidence, the participant lost 10 credits, and had to make ten additional correct identifications to meet the minimum course requirement.

Results Reported in the WLM Article

The principal result was that the 92 participants with more than one year of case work experience made only 61 erroneous comparisons in which they rated their confidence as like that of case work, out of a total of 5,861 latents matched. This is a 1% error rate. A post hoc analysis of the 61 errors found that 59 of the errors could be classified as clerical (for example, mixing up the two hands of the correct suspect). WLM claim that the proper erroneous identification rate was 2 / 5800, or 0.03% (about three errors in ten thousand identifications).

Conclusion Offered by WLM

WLM argue that this fraction of a percent error rate is a more appropriate estimate of an error rate for fingerprint comparisons than any of those extracted from the error rates made on certification tests, on proficiency tests, on other published research on comparison accuracy, or from error estimates based on occasional erroneous identifications discovered during trial or after conviction.

Problems Created by the Switch in Purpose after the Data were Collected

If WLM intended to carry out a study to estimate accuracy of fingerprint comparisons, as reflected in error rates, then the experiment should have insured conditions representative of case work: (1) use of latents varying in difficulty that are typical of those found in case work; (2) use of latents in which the majority do not match the inked exemplars provided; and (3) the participants doing the comparisons are representative of examiners employed in crime laboratories. Further, the experimental design should not permit confounding of procedures: (4) choice of latent difficulty, relative to the examiner, is determined randomly, and not by the instructor or the examiner; (5) course work requirements and accuracy requirements are not confounded; and (6) no hints should be given to assist the examiners once they begin work on a particular latent print. Finally, scoring should reflect unbiased procedures: (7) the response to every latent presented must be scored as either a correct identification, an erroneous identification, a missed identification (the no-response response), or a clerical error; and (8) all scored responses are included in the totals, regardless of confidence.

None of these conditions was met in the experiment reported. In fact, nearly all of them were explicitly violated when the experimental purpose was changed after experiment was designed and the data were collected. We review these violations here.

1. *Difficulty of latents.*

   While the latents included in the packets ranged widely in difficulty, all of the latents were judged to be of value for comparison. The FBI has estimated in a number of Daubert hearings on fingerprint comparison accuracy that over 75% of all latents found in crime scenes are of no value for comparison (because of their poor quality and quantity of detail). Therefore, the latent prints used in this experiment had to be much easier than those found in typical case work. Any estimate of error rate based on these latent prints is confounded with the biased sample of difficulty used in the experiment. These latents were easier than 75% of the latents encountered in case work.

2. *All latents identifiable.*

   Participants were informed that every one of the ten latents in a packet could be identified to one of the inked exemplars. If the exemplars in a packet were selected so that most of them did not match any of the latents, participants would be presented with a greater resemblance to case work, and with a mix of latents in which they had to use their skills to make decisions. It is unclear in which direction the results are skewed by this experimental artifact. Knowing there are no exclusions, a participant might be more willing to guess (and therefore be more likely to make an error); or might persevere longer until he or she finds the correct match (and therefore be less likely to make an error).

3. *Representative skill level of the participants.*

   WLM reported that no certified examiners participated in the experiment. Even among non-certified working examiners, WLM suggested there would be a self-selected bias toward less skilled participants. They also noted that examiners who need more training, but work in under-staffed and under-funded crime laboratories, would be less likely to be allowed to take these courses. So the participants under-represented both the top and the bottom of the skill ladder. These two under-representations are likely to bias the accuracy results, but in opposite directions. There is no way to tease their effects apart, especially in the absence of good biographical data on how much training and experience in latent print comparison each participant had. This lack of representative skill levels means these experimental accuracy results cannot be generalized to working examiners.

4. *Random assignment of latent difficulty to examiner skill.*

   In this experiment, difficulty of the latents was adjusted to the participant's skill. In case work, an examiner rarely has control over the difficulty of the latent he or she compares. This non-random assignment of latents artificially increased accuracy levels.

5. *Course requirements and accuracy requirements.*

   An experimental confounding occurred because participants were free to select the difficulty level of packets so as to insure passing the course. Those who had trouble amassing enough correct identifications (or who made some erroneous identifications), would reasonably opt for easier packets in order to qualify for course completion. While some participants, responding to the training environment of the course, could and did use their free choices to push themselves, and asked for harder and harder latents, they were limited by the realities of the scoring. Consequently, the overall accuracy levels would be artificially increased.

6. *Assistance and hints.*

   The instructor offered hints (naming the perpetrator) whenever a participant was stumped or asked for help. This practice was very helpful in a training environment. However, it is incompatible with the design of an accuracy experiment. As participants, we have no idea of the number of such hints provided, but we know they occurred fairly frequently in the course we attended. The provision of hints was neither scored nor described by WLM. Such hints artificially inflate the percentage of correct identifications.

7. *Not every latent compared and scored.*

   As the experiment was conducted, if a participant found some of the latents in a packet too difficult, or not readily identifiable to any of the inked exemplars, the participant could simply leave the rest of the line blank on the answer sheet. Hence, each participant could omit those latents on which he or she might otherwise have made an error. The participants knew that there was no penalty for leaving the score sheet blank. This procedure artificially reduced the potential for erroneous identifications.

Because every latent is identifiable, when a participant leaves a line on the score sheet blank, the result is equivalent to a missed identification. While this may not enter into court errors, it does indicate that a participant is failing to make an identification when one is possible (the latent is of value, and its donor is in the exemplar packet). Missed identifications are a valid concern in the evaluation of the accuracy of fingerprint comparison work.

8. *Less than case work confidence protected participants from being scored for errors.*

   In this experiment, confidence and scoring of error rate are confounded. Participants knew that if they rated their confidence at case work level and made a mistake, they had to do ten extra identifications. If they rated their confidence lower than case work level and made a mistake, no penalty accrued. While only about 8% of the identifications were made with less than case work confidence, the erroneous identification rate on those was nearly one hundred times that of those made at case work confidence.

Conclusion in this Letter to the Editor

WLM note a number of limitations of their research in their article, some of which have been highlighted above. They do not alert the reader that these limitations fundamentally undermine any conclusions regarding comparison error rates. Based on our analysis, the WLM article in its present form provides no better estimate of the error rates of working examiners than the inadequate data it purports to replace.

Lyn Haber, Ph.D. and Ralph Norman Haber, Ph.D.
Human Factors Consultants
313 Ridge View Drive
Swall Meadows, CA 93514
lhaber@humanfactorsconsultants.com

Authors' Response to Letter:

We (the authors of *A Report of Latent Print Examiner Accuracy During Comparison Training Exercises*) wish to thank the Habers for their letter that raises a number of concerns and criticisms regarding our study of latent print errors in a training environment. We have elected to utilize their abbreviations of authors' names (WLM = Wertheim, Langenburg, and Moenssens respectively, and HH = Ralph and Lyn Haber) for brevity. We are responding to two major areas of criticism in the HH letter: The first issue is the study design and original intent of the study; the second issue is a list of eight limitations that were identified by HH.

Part I – Study Design and Original Intent

With respect to HH's characterization of the study design and intent, much of what they have stated is accurate. The minutiae of the course details and genesis of the project were omitted from the WLM paper for a number of reasons (space, what we believed to be of little interest to the reader, and so forth). If WLM in any way appeared to have misrepresented the facts to the readers, we sincerely apologize—this was not our intent. With HH raising a number of concerns, we will address these concerns.

WLM have issue with several statements of facts by HH (we do not wish to nit-pick and therefore are only addressing the nontrivial differences):

1. Data were collected from 108 participants, not 108 fingerprint examiners as HH stated. Ninety-two of these 108 participants had more than 1 year of experience. We chose to focus primarily on the data of the 92 participants with more than 1 year of experience.

2. The course is not currently, and was not at the time, sponsored by the IAI. No statements have ever been made by any instructor of the course to imply the contrary.

3. The dates of data collection were not as HH stated "from late 2003 and 2004".

4. Students in the course were not required to make a minimum of 60 correct identifications to pass the course (as stated by HH). A student must meet a minimum score of 70 to successfully complete the course. This score is comprised of two metrics: 1 point for each correct individualization, up to a maximum of 60 points, and a 40-point written final examination. Therefore, if a student correctly individualized only 51 latent prints, the student would need to score a 19 (out of 40 possible points) on the final for a minimally passing grade. The reality is that most students do achieve 60 identifications and the final grades tend to be approximately mid- to high-80s for the course.

5. If a student makes an erroneous individualization (at the highest level of confidence), he or she loses 10 points from the maximum of 60 points that can be achieved, and thus the maximum attainable points now for the student is 50. The student, as HH stated, does not have to make up the 10 points by 10 additional correct individualizations. The student cannot make up these lost points. If a student should make 2 or more erroneous individualizations (at high confidence), in all likelihood, he or she will fail the course or can electively be failed by the instructor as well. Additionally in recent courses, a much less severe penalty has been imposed for students who would abuse the confidence rating scheme (by marking all low confidence and therefore be "immune" to making an erroneous individualization). However, the instances observed by the course instructors where serious experts in the field abused such a system are singular or very rare at best.

The remaining descriptions of the course minutiae are relatively accurate and WLM do not dispute these statements by HH.

HH stated that the original purpose of the study was:

> …to use participants from these training groups to develop and validate a measure of latent print difficulty. It had nothing to do with accuracy of comparisons, though assessment of errors was included as one of the measures in the development of quantification of latent print difficulty.

It is true that the initial design of the project, when HH was approached by W, was to develop and validate a measure of latent print difficulty. However, W claims that latent print accuracy was always to be an important aspect of the study and one aspect that critics and examiners would surely have scrutinized, even if not the sole purpose of the study. More importantly, when L was subsequently invited to join the project, after analysis of the data by L, it was clear that the data gathered from 5 or 6 courses were insufficient to perform the intended statistical analysis.

Because there were more than 4000 different latent prints that could potentially be assigned to a course participant and only one participant could work on an exercise packet at a time, after 5 or 6 courses, any particular latent print was rated *at most*, 5 or 6 times by participants. The "medium level" exercise packets tended to be given to participants more often, but generally, the latent prints were examined by only 1 or 2 participants during the periods of data collection. Furthermore, the first initial course data were essentially worthless for this test because the metrics were changed from a 3-point to a 5-point scale. This change was noted in the original WLM article, *but* rather than toss out these data wholesale, because there was an erroneous individualization made at the highest level of confidence in this data set, we elected to keep these data for the analysis of errors.

So while the initial study design had a broader scope, the data were not sufficient for statistical analysis with respect to latent print difficulty. The authors continue to collect these data and still intend to publish such a study. Thus, because the authors continue to gather data in hopes of one day having sufficient data points to quantify latent print difficulty, it seemed tangential, premature, and a waste of premium journal space to discuss the minutiae and genesis of the project. In light of HH's concerns, perhaps we could have given more background information to the reader. Nonetheless, WLM felt that we had sufficient data to provide meaningful estimates of error rates, especially given that no research by the fingerprint community, or the critics, had been offered, yet both parties were providing testimony and opinions, with little data to support these positions.

Which brings us to Part II of HH's letter: how useful are these estimates of error rate given the limitations spelled out by WLM in the original article and the issues raised by HH?

Part II—Additional Study Design Problems Raised by HH

We will address each of HH's eight points.

1. *Use of latents varying in difficulty that are typical of those found in case work*

   HH argue that based on the FBI's research, 75% of latent prints found at crime scenes are of no value for comparison and because all the latent prints in this course have been deemed of value by the instructors, then these latent prints were "easier than 75% of the latents encountered in case work". Firstly, the authors are not aware of any such studies published by the FBI. In discussions with members of the FBI latent print unit and SWGFAST, no one knew of any such study. We are perplexed as to HH's source for this statement. More importantly, whatever the rate of unidentifiable prints from scene work may be, because a latent print is identifiable does not necessarily equate to an "easy" individualization by all examiners. WLM strongly disagree with HH's logic and inferences. A high-quality latent print can quickly become a difficult comparison when the known exemplars are incomplete or of poor quality in the comparative region, or the latent print is not from a completely obvious area of friction ridge skin. Furthermore, whether or not a latent print is "easy" or "difficult" will certainly be a function of many complex variables, including the ability of the examiner [1]. A palm print with more than 20 high-quality minutiae may be very "easy" for someone having taken Ron Smith's "Demystifying Palm Prints", but very "difficult" for someone without that intimate knowledge of palm print comparisons. Finally, and most confusing to WLM, is that if HH were initially interested in partnering with WLM to measure latent print difficulty from data from this course, would not (if based on HH's logic that "identifiable prints are easier than unidentifiable prints") all the prints in this course have been deemed "easy" and a study of latent print difficulty flawed from its inception?

2. *Use of latents in which the majority do not match any of the exemplars*

   WLM addressed this very issue in the original article. We agree that a "perfect error rate study" (hereafter we will refer to as the "Überstudy") should include a significant number of nonmatches. A training course such as this is not an adequate venue to introduce a significant amount of nonmatches.

3. *Participants are representative of examiners practicing in the profession*

   Again, WLM addressed this issue in the original article. We disagree with HH's last statement and conclusion: "This lack of representative skill levels means these experimental accuracy results cannot be generalized to working examiners". HH's statement that these data "*cannot*" be applied to working examiners is a rather strong definitive statement without having sufficient data or research to support that position. WLM simply said that this is a limitation of the study and caution should be applied when using these error rate estimates. We believe that such exclusion of participants may bias the data more toward the "average examiner", and thus these estimates may be a fine estimator for "The Average Latent Print Examiner", but a poor estimate for "The Greatest Latent Print Examiner on the Planet"…and of course… a poor estimate for the "Worst Latent Print Examiner on the Planet". However, this is an unavoidable limitation of performing such a study in this training environment, and this issue could possibly be resolved in the Überstudy.

4. *Choice of latent difficulty should be random*

   WLM agree with HH on this point. This was a limitation we failed to note. It is, however, yet another unavoidable limitation of conducting the study in a training environment.

5. *Course work requirements and accuracy requirements in a training environment confound the study–participants could opt for easier exercises to meet the course requirements*

   WLM recognize again that this is a limitation of performing the study in a training environment; however, this may not be as serious a limitation. Even in casework, examiners do have some control over the difficulty of latent prints that they choose to examine. In some laboratories, experts may never call a latent print "identifiable" that experts in other laboratories would individualize in an instant. This is a function of the variability of examiner ability and average proficiency levels within any given department. Furthermore, if an examiner is faced with a difficult comparison, he or she may have the option to give it to a more experienced examiner to make a determination. So while we recognize this limitation, WLM do not feel it is a critical study limitation.

6. *No hints should be given to assist examiners once they begin their work*

   This is a very valid point. HH stated "we have no idea of the number of such hints provided, but we know they occurred fairly frequently in the course we attended". Quite frankly, WLM do not know exactly how often they occurred either, and WL as instructors both took different approaches to hints. For example, a typical scenario was to examine the latent print (with no knowledge of the correct answer) and say "this is most likely a right thumbprint", using the very clues that are part of the course design. In some instances, it would be narrowed down to a general class of latent print (definitely a "finger" or definitely a "palm"). But WL do agree that in some instances it was narrowed down to a single individual out of 8 possible suspects.

   We have two comments on this issue. First, in our experience, the hints were provided on average maybe once per student, and often resulted in a lower confidence individualization (i.e., using 2 or 1 instead of 3 for highest confidence). A common trend consisted of a student or two, who possessed limited training or experience,

possessed low confidence, or simply of the personality that when stuck would immediately ask for help to avoid frustration. These students would ask for multiple hints while some students simply never asked for assistance and stuck with it. In all, we would estimate that perhaps it evened out to about 1 hint per student. Out of approximately 7500 total comparisons (all levels of confidence), only a relatively small fraction of highest confidence individualizations were affected by this, but an important point nonetheless raised by HH. If our estimate of the frequency of this event is correct, then approximately 100 hints were given to the pool of participants in our study, and even fewer hints resulted in individualizations at the highest level of confidence. This could have affected approximately 1% of the data. However, giving a hint of where to compare will not automatically lead to a correct individualization. *One must remember that finding the match and effecting the individualization are two separate, but sequential, tasks. HH are combining these tasks as a singular event.*

The second comment is that although WLM recognize that this could bias the data, it is not uncommon in case work to consult with colleagues and ask their opinions. Such consultations may or may not bias examiners. We simply do not know how and to what extent, and again, no studies have been provided by critics to show that such interaction will always bias experts. HH finish with "such hints artificially inflate the percentage of correct identifications". HH have no basis for such a definitive statement and although WLM agree that it may have been a biasing contributor, the extent of which is unknown, research or data should be provided by HH to sustain such a definitive conclusion. WLM readily accept that providing hints was yet another limitation of conducting the study in a training environment. We also recognize that a consultation with the course instructor who has access to the "correct answer" is potentially a more serious biasing agent than a consultation in case work as described.

7. *Scoring should be unbiased: the response to every latent presented must be scored as either a correct identification, an erroneous identification, a missed identification, or a clerical error*

HH stated that if a participant chose not to identify a latent print, leaving the answer blank (for which there was no penalty), then "this procedure artificially reduced the potential for erroneous identifications". WLM do not dispute that in the Überstudy this is how the study should be properly set up: scoring correct identification, erroneous identifications, clerical errors, and missed identifications for all latent prints. With such a study design, a false positive error rate (Type I, α factor) and a false negative error rate (Type II, β factor) could be calculated for the ACE-V methodology. Thus TOTAL ERROR = α + β could be calculated and an appropriate testing model and hypothesis rejection criteria proposed for the methodology.

Failure to record an answer in the WLM study does not necessarily reduce the α term as HH have suggested. It would artificially reduce the β term, but we specifically, as have the courts, concerned ourselves with the α term in this study. HH has assumed that if an examiner is forced to evaluate every print that he or she receives, then the examiner will undoubtedly make more erroneous individualizations when attempting comparisons far beyond his or her skill level (in our study, this examiner could simply leave it blank and say "this is just too difficult, please give me another packet"). We propose that examiners, when forced to compare prints that are too difficult, are perhaps more likely to record an opinion of "inconclusive" or "no match" or "insufficient detail to compare". In fact, examiners are trained specifically to not "force the identification" and default to a less conclusive opinion. Based on this belief, we would expect the β term to be significantly higher than the α term and as a profession are generally comfortable with this discordance. In fact, it is a common problem in statistical modeling to find harmony between the α and β terms, because both carry different weights or penalties, depending on the test.

8. *All scored responses are included in the total, regardless of confidence*

   HH stated that "confidence ratings protected participants from being scored from errors". Ironically it was HH who suggested this very scheme to us, for which we are extremely appreciative! WLM disagree wholeheartedly and believe that confidence rating for the individualization is the perfect tool during an experimental environment to evaluate individualizations. This allowed examiners to push themselves to work difficult comparisons, perhaps beyond their comfort or ability level. What was learned from this design was most enlightening: as confidence decreased, more erroneous individualizations were made. WLM fail to see HH's point, especially since both data were reported: total errors and then data filtered separately by level of confidence.

Conclusions

HH stated that WLM "failed to alert the reader that these limitations fundamentally undermine the results of the study". We believe that we were very transparent about the limitations of the study (several of HH's limitations were already identified by WLM) and appreciate the additions identified by HH. WLM feel that these limitations do not "undermine" the results of the study, rather, we leave the reader to weigh the value of these estimators.

HH concluded that the WLM article "provides no better estimate of the error rates of working examiners than the inadequate data it purports to replace". WLM have been transparent about the limitations of the study. But are WLM's data "better" than error rate estimates from anecodotal instances, certification examination pass/fail rates, or CTS testing? It is trivial to argue the value of these data over anecdotes and certification pass/fail rates—the limitations of these sources were addressed in the WLM original article. So we are left with a comparison of our data versus CTS results.

Of all the limitations discussed in the HH letter and the WLM article, CTS data suffer from many of the same limitations, and perhaps even more critical ones. For example, CTS

tests are not taken in a controlled environment. Collaboration may certainly occur during testing and in some agencies, CTS results are reported individually, while in other agencies, CTS results are reported only after peer review by another scientist. CTS does not distinguish between clerical errors and erroneous individualizations, therefore it is reasonable to believe, especially in light of our study data, that CTS error rates are artificially inflated by clerical error rates. CTS data are aggregated to include all levels of experience (and inexperience) and no background information is available for comparison of results to background information.

It has recently been argued by Saks [2, 3] that false positives for fingerprint identification hover around 5% as shown in CTS tests over the last 10 years or so. It is just absurd to believe that erroneous individualizations (false positives) are occurring at such a rate. If we assume a 5% error rate, then in the approximately 200,000 fingerprint cases (reported in 2002) in the U.S. [4], approximately 10,000 erroneous individualizations would be made in a year. Obviously, a percentage of these cases would not bear identifiable prints, but nevertheless, this equates to multiple erroneous individualizations per day! In fact, every AFIS search of a latent print recovered from a crime scene would have a 1 in 20 chance of being erroneously individualized. Because the AFIS databases contain fingerprint records of deceased individuals, elderly individuals, police and public servants, individuals who would have been incarcerated at the time the crime was committed, or individuals who simply could not have committed the crime because of their location, condition, etc., we would just as likely erroneously identify these individuals as the source of the latent print. In other words, there would be a "Brandon Mayfield" every week. Even the most skeptical critic cannot seriously believe this is occurring, and if it were occurring at such alarming rates, we would certainly be aware of it.

Perhaps the greatest strength of the CTS data (compared to WLM's data) is the fact that CTS incorporates nonmatches. As previously stated, future work by WL intend to include nonmatches. This is essential for determining total error.

Finally, although HH have identified eight limitations of the study, only one of these limitations (#7) was applicable to perhaps the most important aspect of this study: the follow-up verification study. Surely, HH must see the value of the inclusion of the verification experiment into the overall study?

WLM strongly feel that although the limitations of the study should caution readers to appropriately weigh these error rate estimates, the study and the data are valuable. The study has produced a number of interesting finds. The research has also shown the community that such a study is not only possible, but also may be very beneficial to the profession. WLM appreciate the open discourse regarding the additional limitations of the study and the weighing of these concerns, but in the end maintain that our study is a better estimator than current sources. It is most certainly better than anecdotal extrapolation and certification examination pass/fail rates, and for the reasons given, more useful and more enlightening than CTS data. How accurate to the true industry false positive error rate, we cannot say, but with further testing and refinement of the study design, we hope to answer this question.

Kasey Wertheim
Glenn Langenburg
Andre Moenssens

References

1. Wertheim, P. The Ability Equation. *J. For. Ident*. **1996**, *46* (2), 149-159.
2. Saks, M., Koehler, J. The Coming Paradigm Shift in Forensic Identification Science. *Science* **2005**, *309* (5736), 892-895.
3. Saks, M., Koehler, J. Response to Langenburg, G. in Questions About Forensic Science. *Science* **2006**, *311* (5761), 607-608.
4. Cole, S. More Than Zero: Accounting for Error in Latent Print Identification. *J. Crim. Law & Criminology* **2005**, *95* (3), 985-1078.