Annex to R.A. Hicklin, et al., In response to Haber and Haber, "Experimental results of fingerprint comparison validity and reliability: A review and critical analysis," *Sci. Justice* (2014), http://dx.doi.org/10.1016/jscijus.2014.06.007.

Detailed list of errors

The publication in *Science and Justice*, "Experimental results of fingerprint comparison validity and reliability: A review and critical analysis"* offers a critique of 13 empirical studies of the performance of latent fingerprint examiners [1-13]†.This table details the errors we found in the Haber and Haber text; these are generally in addition to the issues raised in the previous discussion. These are primarily focused on the portions of their paper that dealt with our Black Box (BB) [1] and Black Box Repeatability and Reproducibility (BBRR) [2] studies; this should not be taken to imply that their commentary on other studies can be assumed to be accurate. Emphasis was added to the Haber and Haber text to indicate errors.

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.

R. Austin Hicklin and Bradford T. Ulery (Noblis)

JoAnn Buscaglia and Maria Antonia Roberts (FBI Laboratory)

Section	Haber and Haber text‡	Rebuttal
Abstract	the experiments did not use fingerprint test items known to be comparable in type and ***especially in difficulty*** to those encountered in casework	85% of BB participants indicated that the overall difficulty of comparisons was similar to their casework; the remainder was equally split (8% easier; 7% harder). The distributions of NFIQ quality metric values for exemplars showed that the exemplars were notably lower quality than operational data [BB SI-1.3].
Sec 2.3	As partially defined in SWGFAST [***16***], we refer to the correct ***definitive*** of exclusion and identification as "***appropriate***," because they reflect a conclusion that matches the ground truth knowledge of the true source of each pair. Conclusions of no-value and inconclusive can be described as "***inappropriate***," because they fail to match ground truth. [16] SWGFAST, ***Standards for the Documentation of Analysis, Comparison, Evaluation and Verification***, 2010. (Latent).	The SWGFAST reference that they cite does not use the term "inappropriate" in this way (neither current [14] nor earlier [15] versions); a 2011 draft for comment of a different document [16] defined inappropriate decisions as nonconsensus decisions (not as Haber and Haber use the term); subsequent versions of that document, in response to public comments, dropped "inappropriate" in favor of "nonconsensus." [17] An individualization or exclusion decision can be regarded as incorrect or erroneous if it contradicts ground truth. However, just because a decision agrees with ground truth, it should not necessarily be considered "appropriate", because there are currently no criteria to determine the sufficiency of information when deciding {value | no-value}, {individualization | inconclusive} {exclusion | inconclusive} - other than consensus among examiners.

* R.N. Haber, L. Haber, Experimental results of fingerprint comparison validity and reliability: A review and critical analysis, Sci. Justice (2014), http://dx.doi.org/10.1016/j.scijus.2013.08.007

† Note that the first 13 references in this paper retain the Haber and Haber reference numbers.

‡ References in this column are the original Haber and Haber reference numbers.

Section	Haber and Haber text‡	Rebuttal
Sec 3.1	The authors tested 169 ***highly trained, highly experienced*** latent print examiners.	The study was not limited to "highly trained, highly experienced" examiners. The examiners had a range of training and experience as described in the Supporting Information. Haber and Haber make a similarly inaccurate statement in the Abstract.
Sec 3.1	***Nearly all*** were certified as ***exceptionally skilled and proficient***, either by the International Association for Identification (IAI), ***the FBI***, or the laboratory in which they worked.	48% were IAI Certified Latent Print Examiners (CLPE). While we understand that some have described this as a test of excellence, that is not the IAI's position. The meaning of agency certifications varies greatly and cannot be taken to mean that an examiner is exceptional; certification or qualification by employers is frequently a requirement for completing training. FBI certification is not mentioned in the BB paper. The FBI does not certify or qualify non-FBI examiners.
Sec 3.2.1	for the same-source pairs for which the correct response was identification, 45% were correctly identified; the remaining 55% were missed identifications. These missed identifications included ***13%*** that were erroneously excluded and ***42%*** that were inconclusive	Only (161+450)/8189 = ***7.5%*** of mated§ comparisons were erroneous exclusions (False negative rate, FNR); (2019+1856)/8189 = ***47.3%*** of mated comparisons were inconclusive [BB, Table S5]. Haber and Haber's incorrect 13% false negative rate bears special mention because a) they repeat the incorrect value five times in the paper, and because b) it appears that they confused prior and posterior probabilities. To make this error, they apparently assumed that FNR was the converse of the negative predictive value (NPV), which is the percentage of exclusion decisions that are true negatives. In BB, with a test mix in which 62% of comparisons were mated pairs, NPV=86.6%. This is a serious misunderstanding: the denominator for NPV is the number of exclusions, whereas the denominator for FNR is the number of mated comparisons. The posterior (PPV and NPV) rates are driven by the mix of mated vs. nonmated data in the test, whereas the prior (FNR and FPR) rates are independent of the test mix. They make an equivalent error regarding Positive Predictive Value (PPV) and False Positive Rate (FPR) in the next section. For a further discussion of prior and posterior probabilities, see the Posterior Probabilities section in BB.
Sec 3.2.1	the remaining 55% were missed identifications. These ***missed identifications*** included 13% that were erroneously excluded and 42% that were inconclusive	"Missed identifications" can be a misleading term and is not used consistently by the latent print community, but is typically restricted to cases where at least one examiner individualizes. The connotation of "miss" is a failure on the part of the examiner who did not individualize. 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.

§ Their use of "same source" corresponds to our use of "mated"; their use of "different source" corresponds to our use of "nonmated."

Section	Haber and Haber text‡	Rebuttal
Sec 3.2.2	When examiners did conclude identification, they were correct ***99.9%*** of the time.	As stated, they are citing the Positive Predictive Value (PPV), which is ***99.8%*** (3661/3669). They are confusing PPV with the converse of the False Positive Rate (FPR=0.1%). The converse of the FPR is not a useful statistic (it would describe "The percentage of mated comparisons that were not erroneous IDs"). They make an equivalent error regarding Negative Predictive Value (NPV) and False Negative Rate (FNR) in section 3.2.1.
Sec 3.2.3	The erroneous exclusion rate (for same-source pairs) was ***13%***	The correct rate is ***7.5%.*** Second mention of this miscalculated percentage.
Sec 3.2.4	***3707*** correct identifications and the ***3949*** correct exclusions	The correct numbers are ***3709*** (40+3669) and ***3947*** (325+3622)
Sec 3.2.5	If the randomly chosen same and different source pairings had been of equal difficulty, the percent of those latent prints rejected as being of no value ***would be equivalent*** for the same- and different-source pairs.	Invalid: see discussion in response to section 4.4, "the different source pairs were easier to compare."
Sec 3.2.5	the results showed that the latent prints that were to have been used in the same-source pairs received ***seven times as many novalue*** conclusions (3389, or 86%) as did the latent prints that were to be used in the different-source pairs (558, or 14%).	The text as written implies prior rates, but the numbers reported are posterior rates. The mated pairs had three times the proportion of no-value latents as did the nonmated pairs (29% of mated pairs vs 10% of nonmated pairs). The numbers they cite are describing the proportion of no-value decisions that were from nonmated vs mated pairs: of the no-value latents, 86% were from mated data vs. 14% from nonmated data. This is a posterior statistic, and therefore the values are dependent on the arbitrary proportions of mates and nonmates.
Sec 3.2.6	The same-source pairs received four times as many inconclusive conclusions (3875, or ***80%***) as did the different-source pairs (1032, or ***20%***).	It appears that they are describing the proportion of inconclusive decisions that were from nonmated pairs (***21%***) vs from mated pairs (***79%***). This is a posterior statistic, and therefore the values are always dependent on the arbitrariness of the mate:nonmate test mix; for example, if prorated so that the mate:nonmate balance is equal, the results would be 30% and 70%. We do not agree that this is a valid method of reporting results: if they want to indicate whether inconclusives were more common in mated or nonmated data, it would be more appropriate to cite the proportion of inconclusives in all presentations of mated and nonmated data (33.5% vs. 18.6%), or in comparisons of mated and nonmated data (47.3% vs. 20.7%). It should be expected that the proportion of inconclusives in mated data will be higher than in nonmated data: this cannot be assumed to be an effect of differing difficulty.
Sec 3.2.7	This result ***is not reported*** in Ulery et al.	The data is plainly presented in the Results section of the paper as Figure 2 (a mosaic plot showing determinations by mating) and again in tabular form in the SI (Table S5, from which they got these numbers).

Section	Haber and Haber text‡	Rebuttal
Sec 3.3	The average number of examiners that viewed each pair was ***39*** examiners (23%).	Each image pair was examined by an average of ***23*** participants.
Sec 3.3.1-3.3.3	Only 43% of the value conclusions for the latent prints were ***unanimous*** […] Only 15% of the same-source pairs were identified ***by all of the examiners***. […] Seventy-five percent of the different-source pairs were correctly excluded ***by all examiners***	Their use of unanimity as a measure of reliability is problematic, as it is highly dependent on the number of examiners tested: unanimity among two examiners is quite different (and much more likely) than unanimity among a mean of 23 examiners (the BB values they cite here); for this reason, measurements of unanimity cannot be expected to be comparable across studies. In BBRR we use percentage agreement to report inter-examiner reproducibility, which avoids these issues.
Sec 3.3.2	Only ***15%*** of the same-source pairs were identified by all of the examiners. Of the remaining 85%, ***46%*** of the pairs were unanimously inappropriate or ***erroneous (exclusion) conclusions.***	***10%*** of mated pairs were unanimously individualized. (BB Fig 5) For ***47%*** of the mated pairs, none of the examiners individualized: all determinations for these mated pairs were no-value, inconclusive, or erroneous exclusions. 38% of mated pairs were unanimously inconclusive or no value. No mated pairs were unanimous erroneous exclusions.
Sec 3.3.3	***Seventy-five percent*** of the different-source pairs were correctly excluded by all examiners	The correct number is ***27%*** (61/224). (Exact value was not reported in BB, but can be estimated from BB Fig. 5 marginal red histogram, or BB Fig S6. The exact value can be calculated easily using the BB data provided to Haber and Haber**.) Note that while their section is entitled "Consensus on exclusion conclusions," here they are discussing consensus on nonmate data instead of exclusions.
Sec 3.3.3	***20%*** received differing conclusions across the examiners.	The correct number is ***69%*** (154/224). (Exact value was not reported in BB, but can be estimated from BB Fig S6. The exact value can be calculated easily using the BB data made available to Haber and Haber.)
Sec 3.4	[re Reliability of examiners] These data ***were not reported,*** but we used the authors' Fig. 7 for estimates.	In [Haber and Haber Sec 2.5] "reliability of examiners" is defined as "the percentage of examiners who agreed with one another on the responses they gave." Based on this definition, they are describing interexaminer reproducibility of decisions. This is the topic of the BBRR paper that they claim to have reviewed: BBRR reported intraexaminer results from the follow-on repeatability test, as well as additional interexaminer reproducibility results from the initial BB test. In [Haber and Haber Sec 3.4], they are describing the distributions of responses among examiners, which is not a relevant proxy for interexaminer reproducibility results. In addition, they made multiple errors in reporting the results of distributions of responses among examiners: the caption for the figure they cite points to Table S7, which contains the values in question.

** Haber and Haber requested and received a copy of the test and survey results from the "Black Box" study, which are available at http://www.fbi.gov/about-us/lab/scientific-analysis/counterterrorism-forensic-science-research/black-box-study

Section	Haber and Haber text‡	Rebuttal
Sec 3.4	The average identification rate for the same-source pairs was ***45%.***	The correct number is ***32%*** (mean TPR_{PRES} for all presentations, Table S7), but for individualization determinations, a more appropriate result would be mean TPR for all comparisons in which the examiner determined the latent was of value for individualization (mean TPR_{VID}), or 61%.
Sec 3.4	The best subject identified ***65%*** of the same-source pairs presented	The correct number is ***57%*** (max TPR_{PRES}, Table S7).
Sec 3.4	the poorest identified only ***20%***	The correct number is ***11%.*** (min TPR_{PRES}, Table S7)
Sec 3.4	On the different-source pairs, the average correct exclusion rate was ***79%.***	The correct number is ***71%.*** (mean TNR_{PRES}, Table S7)
Sec 3.4	the poorest excluded only ***40%*** of the different-source pairs.	The correct number is ***7%.*** (min TNR_{PRES}, Table S7)
Sec 3.5	This was ***strictly a reliability-of-conclusions study***: how many times would the same conclusion be given ***when a trial was repeated a second time*** at some later date.	BBRR reported not only intraexaminer repeatability results, but also additional interexaminer reproducibility results from the initial test (BB). Despite the title of the paper, Haber and Haber fail to recognize that BBRR contained reproducibility data, which they say were "not reported" (Sec 3.4).
Sec 3.5	The authors ***were not able to measure*** the reliability of examiners' consensus among one another, because each examiner was given different pairs in the repeatability testing	Interexaminer reproducibility of decisions (which Haber and Haber call "reliability of examiners") was measured on the data from the initial assignment of the 100 pairs, not on the results from the second assignment. Reproducibility is discussed and extensively reported throughout BBRR.
Sec 3.6	Overall, ***about 90%*** of the repeated test items received the same response on their second presentation. […] The Ulery et al. [2] result is that ***10%*** of the conclusions were inconsistent within the same examiners.	In BBRR we report many such results, but 90.3% and 85.9% (BBRR Fig 6, 3-way mates and nonmates) are the appropriate values. At a minimum, the caveats "overall" and "about" should be included when restating the converse 10% value.
Sec 3.6	On the 16 same-source pairs, 89% of the original identification conclusions were repeated, and 11% were changed, most to inconclusive, and ***a few to no-value***.	***None*** were changed to no-value.
Sec 3.6	On the different-source pairs, ***90%*** of the exclusions were repeated	The correct value is ***91%*** (90.6%).
Sec 4.1	The examiners were asked on a post-experiment questionnaire if they used the conclusion "of-value-only-for-exclusion" in their normal casework. ***Only 17%*** said yes. The ***remaining 83%*** of the examiners may have interpreted this ***unfamiliar*** conclusion in a variety of different ways [...]	It is not accurate to say it was "unfamiliar" to 83% of the BB examiners: 30% of BB participants use the conclusion (17% in standard practice, 13% used only on request). Of the BB participants, 55% consider VEO prints not of value, and 14% consider them of value.

Section	Haber and Haber text‡	Rebuttal
Sec 4.1	When examiners judged a latent to be of-value-only-for-exclusion, they were still allowed to compare it and then offer conclusions ***inconsistent with the "only" in the conclusion: they were allowed to conclude*** identification or ***inconclusive.***	Inconclusive determinations are consistent with VEO value assessments. For VEO, participants were instructed (BB SI-1.5): "Value for exclusion; only applies if 'Not of value for individualization'; The impression contains some friction ridge information (level 1 and/or level 2) that may be appropriate for exclusion if an appropriate exemplar is available." They continue this misconception when they suggest "better and more useful scoring" in which a VEO latent cannot be permitted to result in an inconclusive.
Sec 4.1	The value judgment of "of value only for exclusion" ***was then re-recorded as "of value."***	False. We report separately categories such as VEO individualizations and VID individualizations (e.g., Fig. 7). Depending on the analysis, we sometime aggregate results to address specific questions or to draw attention to broad patterns in the data.
Sec 4.1	After the exemplar appeared, ***only 500*** of those 3122 exemplar prints (16%) were actually excluded as the source.	The correct number is ***486***. (BB Table S5)
Sec 4.1	Nearly all of the rest (2622) were judged inconclusive (84%),	The correct number is ***2596***. (BB Table S5)
Sec 4.1	All trials labeled as of-value-only-for-exclusion that were not excluded ***were combined with value-for-comparison*** when scored	In BB, Value for comparison (VCMP) "includes comparisons where the latent was of value for exclusion only (VEO) as well as VID." VEO is not combined with VCMP; it is part of VCMP. To say the data was combined when scoring is misleading: summary statistics were reported in multiple ways and a detailed breakout of the counts was provided in the SI (uncombined).
Sec 4.2	The ***best design*** would have been to present the same 100 pairs to each subject.	That design would have conflicted with our objectives and limited the value of the study. We deliberately (and appropriately) made an early design trade-off decision, in which we chose to sample a greater variety of image pairs to better understand the effect of the fingerprints on variance. In BB, we stated "The number of fingerprint pairs used in the study, and the number of examiners assigned to each pair, were selected as a balance between competing research priorities: Measuring consensus and variability among examiners required multiple examiners for each image pair, while incorporating a broad range of fingerprints for measuring image-specific effects required a large number of images." To be more explicit: we deliberately selected a range of data (including difficult nonmated data) so that we could assess how a variety of fingerprint attributes would affect examiners' decisions. If we had used the same image pairs for each examiner, the results would have been based on a much more limited pool of fingerprints: it would have allowed us to compare these examiners better, but given us less information about the effects of a broad variety of latent prints, which was a major goal of the study.
Sec 4.4	The examiners made more accurate appropriate conclusions for the different-source pairs (71%) as compared to the same-source pairs ***(24%).***	Since "accurate appropriate" is never clearly defined, it is difficult to determine what they are reporting. Since for nonmates they appear to use TNR_{PRES} (true negative rate for all presentations = 71.2%), the corresponding value would be TPR_{PRES} = ***32.0%***.

Section	Haber and Haber text‡	Rebuttal
Sec 4.4	The examiners judged only one-fourth as many of the different-source pairs inconclusive ***(20%)*** as compared to inconclusive conclusions for the same-source pairs ***(80%).***	The correct numbers are ***21%*** and ***79%***. We do not agree that this is a valid method of reporting results: see discussion in 3.2.6
Sec 4.4	erroneous exclusions among the same-source pairs (***13%***).	The correct number is ***7.5%***. Third mention of this miscalculated percentage.

Section	Haber and Haber text‡	Rebuttal
Sec 4.4	the different source pairs ***were easier to compare*** than the same-source pairs, so poorer performance would be expected on the more difficult same source pairs. There is one finding that can only be interpreted as a difference in difficulty. The examiners made only one-seventh as many no-value conclusions among the latent prints intended for the different-source pairs (14%) compared to the latent prints intended for the same-source pairs (86%). Since the no-value conclusions were made before the latent was paired with an exemplar, the latent prints in the pool for different-source pairs ***must have been significantly easier***. […]The authors stated that their selection procedures were specifically designed to make the different-source pairs more difficult to compare than the same-source pairs. However, these results suggest that ***this manipulation failed***, and ***the different-source pairs were actually easier***.	Haber and Haber use the term "difficult" differently than we do: the differences need to be understood to understand this discussion. Our use of "difficulty" is based on each examiner's assessment regarding how easy or difficult it was to make a determination: difficulty in comparison may be seen as how close the comparison is to the line between two conclusions. Haber and Haber generally use "difficulty" to refer to how easily an examiner can correctly determine that the images are mated or nonmated. The differences in definition are most notable for very poor-quality prints: Haber and Haber would consider a nearly-blank print to be difficult. Using our definition these are not difficult: it would be very easy for an examiner to determine that a nearly-blank print is an inconclusive. Data selection for the mated and nonmated image pairs was performed separately in order to focus on the most challenging nonmate comparisons we could select. We began with the same pool of latents, but the process of selecting the nonmates filtered out many of the no value prints because comparisons with such prints are not challenging: including more prints of no value would have resulted in easy inconclusive decisions, and would not have made the nonmated data more difficult. Data selection for mated pairs was based on arbitrarily selecting a mated exemplar for each latent, which resulted in a large proportion of latents that were subsequently assessed by some or all examiners as no value, and a random distribution of difficulty for the resulting comparisons. As described in BB (SI 1.3): "Approximately one-half of the non-mated pairs were selected by an experienced latent print examiner, who was not a participant, using one of two processes with the objective of maximizing the difficulty of comparisons: either the examiner selected from the twenty candidates returned by IAFIS the exemplar that would result in the most difficult comparison (18%), or the examiner selected an exemplar from the neighboring fingers from the correct subject (29%). For the remainder of the non-mated pairs, the first exemplar in the list of IAFIS candidates was selected. The process of selecting challenging non-mated pairs was time-consuming and therefore was not pursued for latents that were considered to be of no value by the subject matter experts doing data selection; as a result of this, the latents in mated pairs included a greater proportion of poor-quality latents than did the non-mated pairs." Haber and Haber's statement that our data selection process was "manipulation" is not appropriate: their misunderstanding does not mean that it failed. See also the discussion of difficulty in response to the Abstract.

Section	Haber and Haber text‡	Rebuttal
Sec 4.4	The examiners made only ***one-seventh*** as many no-value conclusions among the latent prints intended for the different-source pairs (***14%***) compared to the latent prints intended for the same-source pairs (***86%***).	We do not agree that their "one-seventh" is an appropriate method of reporting results: see discussion in response to Section 3.2.5.
Sec 4.4	the authors ***do not report*** the repeatability scores separately for the easy versus difficult pairs	We reported repeatability by comparison difficulty (BBRR Fig 4, Table 4, and Fig S7).
Sec 4.8	The uncontrolled difference in difficulty between the same- and different-source pairs suggests that the low erroneous identification rate found was ***due to easy*** different-source pairings, and ***would have been higher*** had comparable pairings been used.	False: see discussion in response to Section 4.4, "the different source pairs were easier to compare."
Sec 4.8	the nonrandom sampling of the examiners who served as subjects ***suggests*** that the erroneous identification rate would have been higher among average examiners.	The fact that the errors were concentrated on a specific highly complex combination of processing and substrate suggests that the rate might have more to do with the sample fingerprints than with the sample participants. Given the rarity of erroneous identifications, they might be associated with a few specific examiners, in which case the group average might be a poor predictor.
Sec 4.8	Further, in spite of the authors' intentions to respond to the National Academy of Sciences [18] critiques, these results do not provide evidence of the validity or reliability of the ACE method, since that method explicitly was not assessed in these studies.	The "ACE method" was not the subject of our study. Our work responds to NAS Recommendations 1(a)(b)(c), 3(a)(b) and 8 by increasing our understanding of "the accuracy of examiner conclusions, the level of consensus among examiners on decisions, and how the quantity and quality of image features relate to these outcomes." [BB] No single research study will adequately address any one of the NAS recommendations.
Sec 4.8	The reliability results suggest that the outcome of a particular comparison ***depends more*** on which examiner is assigned to the case than on the physical characteristics of the stimulus print to be compared.	This claim is baseless: there is no data in BB or BBRR to suggest that the examiners' reproducibility and/or repeatability outweigh the effect of the print being compared; physical characteristics were not measured.
Sec 5.5.1	Across the five experiments, correct identification of the same source pairs ranged from very high accuracy of 91% (Langenburg [5]), to a low of 45% (Ulery et al. [1]). Correct exclusion in the different source pairs ***ranged from 79% (Ulery et al. [1]) to 21% (Langenburg, [5]).***	These results are not at all comparable. In our study, "exclusion" refers to the exclusion of a single finger. In the Langenburg study [5], each latent was compared to a set of ten-print cards from eight "suspects"; all eight subjects had to be excluded to count as the ACE process resulting in an exclusion. The Langenburg study had only six participants, all from one organization, two of whom were not yet certified. This is but one example of the substantial differences among studies that Haber and Haber fail to acknowledge.
Sec 5.5.2	The erroneous identification rate ranged from a low of 0.1% (Ulery et al. [***11***])	Reference should be [1]

Section	Haber and Haber text‡	Rebuttal
Sec 5.5.2	The erroneous exclusion rate ranged from a low of 1% (Langenburg [5]) to a high of ***13%*** (Ulery et al. [1]).	The correct number is ***7.5%***. Fourth mention of this miscalculated percentage.
Sec 7	five limitations that stem from ***poor reliability of the results***	Haber and Haber fail to account for a critical factor that undermines their conclusion: the measures vary from study to study because the studies are not measuring the same thing. It is ironic that Haber and Haber are in one section (Sec 7.2) criticizing the studies for not including statistical tests, but here are willing to make comparisons and draw conclusions that necessarily require such information (e.g. confidence bounds).
Sec 7.1	In addition, most of these experiments demonstrated that the amount of agreement between examiners in their conclusions was low. Except for conditions where the results reach a ceiling at close to 100%, the examiners ***rarely reached a unanimous conclusion*** for the pairs they compared.	This is a tautology: restated, this sentence would read, "Except when examiners were unanimous, examiners rarely reached unanimous conclusions."
Sec 7.1	The low reliability in the experimental results precludes inference of performance levels in casework	In section 7.1, they are using "reliability" in two different ways: the heading and the first paragraph refer to the variability of measurement results across experiments, but then the second paragraph changes to the very different topic of inter-examiner reproducibility. Inter-examiner reproducibility in no way would preclude inference to casework in general: examiners in operational casework may well show the same imperfect rates of reproducibility as shown in these studies. If they are referring to inference to a specific case (e.g. predicting the performance of an individual examiner who might be testifying), such inference is indeed problematic, but not just because of the low reproducibility found in these studies, but because inference from general results to specifics is always a concern.
Sec 7.1	Ulery et al. [2] found that ***10%*** of the conclusions reached by examiners changed when the same pairs were compared a second time.	Misleading; see comment on section 3.6 ("Overall, about 90% of the repeated…").
Sec 7.1	These within-subject results suggest that the performance of an individual examiner at a given time ***does not predict*** the same examiner's performance at a later time.	It not only predicts, but is correct about 90% of the time for mated data, or 86% for nonmated data.

Section	Haber and Haber text‡	Rebuttal
Sec 7.2	***Only two*** of these 13 experiments [10,13] were published in rigorously vetted scientific journals.	False. In fact, ***11*** of the 13 articles appear in peer-reviewed journals: all except [6 and 12]. BB [1], which received the majority of Haber and Haber's criticism, was published in the *Proceedings of the National Academy of Sciences*, one of the world's most highly-regarded scientific journals; Haber and Haber omitted the name of the journal from the references. The editor for BB was Dr. Stephen Fienberg, a leading US statistician. In the Acknowledgments, Haber and Haber state "We presented many of the results of this article during a Frye Hearing (Illinois v. Robert Morris) in May, 2012, in which we had been retained as defense experts." This error is a continuation of Ralph Haber's statements in that hearing, in which he erroneously testified [18] that "Some of them are published without any reviews or an absolute minimum review. The Ulery study -- Both Ulery studies were published that way. They were published in a journal in which they -- the FBI had to pay to get the journal to publish it rather than it being accepted because it was a good experiment."
Sec 7.2	[Statistical] tests that ***cannot be performed*** when the results cluster at perfect performance.	False: statistical tests can be performed (and confidence intervals measured) even when 100% of responses are the same.
Sec 7.3	***no differences*** between supposedly easy and difficult prints were found [3,13]	BBRR reports the opposite (BBRR Fig 7 and Fig S7).
Sec 7.3	the differences between same- vs. different source pairs in Ulery et al. [1] ***were opposite to the authors' intent***	Incorrect: see discussion in this table under Haber and Haber Sec 4.4 ("the different source pairs were easier to compare…") for an explanation of Haber and Haber's invalid rationale. See also the discussion of difficulty in response to the Abstract.
Sec 7.3	The same is true for the ***difficulty of exemplar prints***	The NIST fingerprint image quality algorithm (NFIQ) has been widely used for this purpose for almost 10 years.
Sec 7.5	The no-value conclusion was estimated to occur for between 50% and 75% of all latent prints brought to an examiner. When comparisons are made to the remaining latent prints, virtually all are exclusions. ***Inconclusive judgments are rare in casework***, and identification conclusions are even rarer: estimated to be less than 1% of all conclusions reached in casework.	These estimates are problematic, because rates vary dramatically by agency and case type. Agency policies can affect decision rates: for example, in the BB participant survey (BB SI-1.4), 32% of participants were not permitted to make inconclusive determinations and an additional 19% were discouraged from making inconclusive determinations; 23% never used exclusion as a determination; the rates of inconclusive decisions will certainly be lower for agencies in which examiners are discouraged from making inconclusive decisions. Collection procedures may affect rates: crime scene investigators may filter out most no-value prints before they are provided to examiners. The type of case will also affect conclusion rates: for major crimes, more latents of marginal value may be collected; for minor crimes, crime scene investigators may not bother collecting latent prints, or collect only the highest-quality prints; cases with suspects are far more likely to be same-source than AFIS searches and therefore will have a far greater proportion of individualizations.

Section	Haber and Haber text‡	Rebuttal
Sec 7.5	If the purpose of these experiments, even in part, is to estimate the erroneous identification rate, then the experiments should contain a substantial number of different-source pairs, because only different-source pairs can provide an opportunity to make erroneous identifications. These experiments included ***relatively few*** different source pairs: the overall average was only about a quarter of the pairs. [...]. If the purpose of the experiments is to estimate erroneous identification rates, the ***prevalence*** of same-source over different-source pairs is bad science and is a biased experimental design.	Faulty logic: the erroneous ID rate (FPR) is not affected by the proportions of mates and nonmates. Although they start the paragraph accurately stating that "the experiments should contain a substantial number of different-source pairs," the rest of the paragraph misleadingly digresses into the proportions. The relative proportions are critical to the posterior rates (PPV and NPV), but these proportions are not known in casework and may be expected to vary substantially. An appropriate approach is to chart the posterior rates as a function of the variation in such proportions, as we did in BB (Fig 4). We collected 5,543 responses on nonmated pairs and observed 6 erroneous individualizations among 4985 comparisons. A larger sample size would have additional precision, but as Haber and Haber point out, the more important issue is how well our sample reflects a population of interest.
Sec 7.6	suggest that they were ***substantially above average*** in skill and experience	Incorrect. See response to section 3.1.
Sec 7.7	In casework, the minimum-sized case consists of a single latent print and ***ten exemplar prints from a single suspect***.	AFIS casework is frequently based on the comparison of one latent and one exemplar, which corresponds directly to the scenario used in the BB and BBRR studies. Agencies' standard operating procedures (SOPs) vary in particular in the types and implications of value and inconclusive decisions. Since the latent examiners represented a broad spectrum of the community, a uniform approach based on processes used by most latent examiners was used. Note also that both AFIS and non-AFIS casework may involve comparisons of a latent to another latent.
Sec 7.9	Substantial research has shown that subjects who know they are being tested ***perform better*** than when the tests are not announced and cannot be differentiated from routine work (e.g., Koppl et al. [28]).	While participants in tests may indeed have different performance than in routine work, it is not reasonable to conclude that the results are necessarily better in the tests: a few examiners who are not taking the test seriously could have notably affected the results of a study, especially with respect to rare events. For example, we do not know if the examiner who made two erroneous individualizations was acting as s/he would have in routine work, or was just tired and apathetic, given it was just a test. It seems likely that at least some of the participants took the test less seriously than casework, given the serious implications of actual casework, and the absence of any negative implications on an anonymous test.

Section	Haber and Haber text‡	Rebuttal
Sec 7.10	***Consequently, casework examinations are, overall, more difficult than the comparisons in these experiments,*** and the ***accuracy and reliability results of these experiments are inflated*** compared to casework.	Two errors: 1) It is not reasonable to assume that complete AFIS-generated candidate lists would necessarily be more difficult than single selections: the AFIS ranks candidates in decreasing order of similarity, and so overwhelmingly the first candidate is the most similar. It is unusual for lower-rank candidates to be nearly as challenging as higher-rank candidates. 2) If full candidate lists were included, the number of easy exclusions would have increased, so while the number of false positives would not be expected to increase (numerator of false positive rate), the number of nonmate comparisons would have increased by 20 (denominator of false positive rate), so the accuracy as measured in the test would decrease very substantially. Note also that the size of the AFIS database is the primary factor in selecting similar nonmates, and results from AFIS of different size should not be equated. BB selected nonmates from the FBI's IAFIS, which at the time contained 58 million subjects (580 million distinct fingers).
Sec 10	On the assumption that these experiments reflect real life, and that ***every same-source pair is from a "guilty" person*** and ***every different-source pair is from an "innocent" person,***	In "real life", same-source pairs are very frequently "elimination prints" from the victim(s), law enforcement, or other people with legitimate access.
Sec 10	to a high of ***13%*** (Ulery et al. [1]).	The correct number is ***7.5%***. Fifth mention of this miscalculated percentage.
Sec 10	If these data could be generalized to casework, they would indicate that a very large number of "guilty" perpetrators remain at large to commit further crimes. […] If the results from these experiments were generalizable to casework, fingerprint comparison evidence ***would leave guilty perpetrators free***	Haber and Haber's usage of "missed identifications" was discussed above in the response to 3.2.1. Inconclusives indicate the lack of evidence (when examiners agree) or debatable evidence (when examiners disagree). It is troubling that Haber and Haber are suggesting that conclusions should be made even in these cases.
Sec 11	***exemplar print difficulty***	Repeat of previous error: the NIST fingerprint image quality algorithm (NFIQ) has been widely used for this purpose for almost 10 years.
References	[1] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, www.pnas.org/cgi/doi/10.1073/pnas.1018707108.	No reference to the journal. Should be Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2011) Accuracy and reliability of forensic latent fingerprint decisions. ***Proc Natl Acad Sci USA 108(19): 7733-7738***. (www.pnas.org/cgi/doi/10.1073/pnas.1018707108)
References	[2] B.T.Ulery, R.A.Hicklin, J. Buscaglia,M.A. Roberts, (2012), Repeatability and reproducibility, of decisions by latent print examiners, ***www.plosone.org/article/info:doi/10.1371/journalpone 0032800***.	No reference to the journal, and invalid URL. Should be Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2012), Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. ***PLoS ONE*** 7:3. (***http://www.plosone.org/article/info:doi/10.1371/journal.pone.0032800***)

Section	Haber and Haber text‡	Rebuttal
References	[4] Z.W. Evett, R.L. Williams, Review of the 16 point fingerprint standard in England and Wales, Forensic Science International 46 (1996) 49–73.	Should be [4] Z.W. Evett, R.L. Williams, Review of the 16 point fingerprint standard in England and Wales, ***J. Forensic Identification*** 46***:1*** (1996) 49–73.
References	[6] S.B. Meagher, Report of the Federal Bureau of ***Identification's*** Mitchell Survey, US v. Mitchell, 365 F, ***1998***, (3d. Circuit).	Federal Bureau of ***Investigation***, ***1999***. A more complete reference would be 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)*
References	[7] I.E. Dror, D. Charlton, A. Peron, Contextual information renders experts vulnerable to making erroneous identifications, Forensic Science International ***56*** (2006) 74–78.	I.E. Dror, D. Charlton, A. Peron (2006) Contextual information renders experts vulnerable to making erroneous identifications. Forensic Science International ***156***(1):74–78.
References	[12] S. Gutowski, Error Rates in Fingerprint Examinations: The View in 2006, ***Forensic Science*** Bulletin, *Autumn*, 2006. (2006).	[…]***The Forensic*** Bulletin, Autumn 2006, ***18-19***.

Detailed list of errors in the initial 2013 publication

A previous version of the Haber and Haber article was published online 18 November 2013.[††] The revised 2014 article corrects some of the errors in that original article. Because that 2013 version was publicly distributed, we include in this table details of the errors we found in the 2013 Haber and Haber text that were corrected in the 2014 revision.

Section	Haber and Haber text[‡‡]	Rebuttal
Sec 2.3	SWGFAST [16] refers to the definitive conclusions of exclusion and identification as ***"appropriate,"*** [...] SWGFAST defines conclusions of no-value and inconclusive as ***"inappropriate."*** [16] SWGFAST, ***Standards for the Documentation of Analysis, Comparison, Evaluation and Verification***, 2010. (Latent).	This statement contains multiple errors. The term "appropriate" was used in BB prior to its being proposed by SWGFAST: "The best information we have to evaluate the appropriateness of reaching a conclusion is the collective judgments of the experts." [1] Therefore, the term as used in [1,2, and 17] refers to nonconsensus decisions, either inappropriately inconclusive (cautious), or inappropriately conclusive (incautious). It does **not** apply to **all** inconclusive or no-value decisions: for example, it is entirely appropriate for a consensus of examiners to assess the poorest quality prints as having no value for comparison. The SWGFAST reference that they cite does not use the term "inappropriate" in this way (neither current [15] nor earlier [16] versions); a 2011 draft for comment of a different document [17] defined inappropriate decisions as nonconsensus decisions (not as Haber and Haber use the term); subsequent versions of that document, in response to public comments, dropped "inappropriate" in favor of "nonconsensus." [18] (R. Austin Hicklin was one of the authors of the original document for SWGFAST.) Haber and Haber's misuse of "appropriate" is especially problematic because they proceed to misuse the term 17 times in reviewing other papers, and this aggravates their underlying misunderstanding of the role of inconclusive and value determinations.
Sec 2.3	***None of the experiments*** in the corpus reported their results in this way. [regarding use of "inappropriate"]	The term "appropriate" was initially used in BB prior to its being proposed by SWGFAST. We did not, however, report results using Haber and Haber's misreading of the terminology.
Sec 3.1	***returned the disk to the authors with their conclusions*** when finished	No such procedure was followed. The disks were not rewritable; responses were returned via email.
Sec 3.4	This range from 65% to 20% shows that examiners were responding with a different distribution of conclusions, ***indicating low agreement among examiners***.	Faulty analysis: examiners were not assigned the same image pairs; this statistic has nothing to do with interexaminer agreement on the same image pairs. Haber and Haber's review entirely overlooks our analysis of interexaminer agreement on this data [BBRR].

[††] R.N. Haber, L. Haber, Experimental results of fingerprint comparison validity and reliability: A review and critical analysis, Sci. Justice (***2013***), http://dx.doi.org/10.1016/j.scijus.2013.08.007

[‡‡] References in this column are the original Haber and Haber reference numbers.

Section	Haber and Haber text‡‡	Rebuttal
Sec 3.5	If an examiner had originally made an erroneous identification ***and/or a missed identification,*** those were re-presented	Erroneous individualizations and erroneous exclusions were reassigned for the repeatability test; "missed identifications" were not reassigned. See also discussion under section 3.2.1, "the remaining 55% were missed identifications…"
Sec 4.3	While the authors ***do not describe how many of the same latent prints*** and the same exemplar prints were presented more than once to an examiner,	BBRR delineates the 900 cases in which an examiner saw the same latent twice.
Sec 4.3	To make 744 pairs, each latent print ***had to have been used at least twice*** for each subject	Each latent print was used at least twice to construct the pool of 744 pairs - but each participant was assigned only 100 image pairs and did not see each latent print at least twice. BBRR states that an examiner saw the same latent twice in 900 out of the 17,121 presentations.
Sec 7.3	***Two experiments*** used AFIS similarity ratings to select similar pairings (Langenburg et al. [3] and Tangen et al. [13]).	BB also used AFIS to select similar different source pairs, as discussed in the response to section 4.4 ("the different source pairs were easier to compare…").
Sec 7.10	None of the experiments presented AFIS-produced candidate exemplars to be compared to the latent prints. ***(Langenburg et al. [3] and Tangen et al. [13]*** used AFIS to select prints for use in their experiments, but their subjects never saw more than one AFIS candidate in the experiments.)	BB also used an AFIS to select similar nonmates.
References	[1] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, ***www.pnas.org/cgi/doi/10.1073/pnas.101 87071082011.***	No reference to the journal, and invalid URL. Should be Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2011) Accuracy and reliability of forensic latent fingerprint decisions. ***Proc Natl Acad Sci USA 108(19): 7733-7738***. (***http://www.pnas.org/content/108/19/7733.full.pdf***)
References	[2] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Repeatability and reproducibility, of decisions by latent print examiners, ***PhoS 7*** (2012), ***http://dx.doi.org/10.1371/journal.poon.0032800.***	Incorrect journal, and invalid URL. Should be Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2012), Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. ***PLoS ONE*** 7:3. (***http://www.plosone.org/article/info:doi/10.1371/journal.pone.0032800***)
References	[7] I.E. Dror, D. Charlton, A. Peron, Contextual information renders experts vulnerable to making erroneous identifications, ***J. Forensic Identification 56*** (2006) 74–78.	I.E. Dror, D. Charlton, A. Peron (2006) Contextual information renders experts vulnerable to making erroneous identifications. ***Forensic Sci Int. 156***(1):74–78.
References	[9] L.J. Hall, L. Player, ***With*** the introduction of an emotional context affect fingerprint analysis and decision-making, Forensic Sci. Int. 181 (2008) 36–39.	***Will*** the introduction of an emotional context affect fingerprint analysis and decision-making***?***

Section	Haber and Haber text‡‡	Rebuttal
References	[10] G. Langenburg, C. Champod, P. Wertheim, Testing for potential contextual bias effects during the verification stage of ACE-V methodology when conducting fingerprint comparisons, J. Forensic Sci. ***50*** (2009) ***1–12***.	G. Langenburg, C. Champod, P. Wertheim (2009) Testing for potential contextual bias effects during the verification stage of ***the*** ACE-V methodology when conducting fingerprint comparisons, J. Forensic Sci. ***54***(3):***571-582***.

References

1 B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts (2011) Accuracy and reliability of forensic latent fingerprint decisions. *Proc Natl Acad Sci USA*, 108(19):7733-7738. (http://www.pnas.org/content/108/19/7733.full.pdf)
2 B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts (2012), Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. *PLoS ONE* 7:3. (http://www.plosone.org/article/info:doi/10.1371/journal.pone.0032800)
3 G. Langenburg, C. Champod, T. Genessay (2012) Informing the judgments of fingerprint analysts using quality metric and statistical assessment tools. *Forensic Sci Int*. 219(1):183-198. (http://dx.doi.org/10.1016/j.forsciint.2011.12.017)
4 Z.W. Evett, R.L. Williams (1996) Review of the 16 point fingerprint standard in England and Wales. *J. Forensic Identification* 46(1):49–73.
5 G. Langenburg (2009) 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 Identification* 59(2):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 (2006) Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Sci Int.* 156(1):74–78.
8 I.E. Dror, D. Charlton (2006) Why experts make mistakes. *J. Forensic Identification* 56(4):600–616.
9 L.J. Hall, L. Player (2008) Will the introduction of an emotional context affect fingerprint analysis and decision-making? *Forensic Sci. Int.* 181(1):36–39.
10 G. Langenburg, C. Champod, P. Wertheim (2009) Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons, *J. Forensic Sci.* 54(3):571-582.
11 K. Wertheim, G. Langenburg, A.A. Moenssens (2006) A report of latent print examiner accuracy during training exercises. *J. Forensic Identification* 56(1):55-93.
12 S. Gutowski (2006) Error Rates in Fingerprint Examinations: The View in 2006, *The Forensic Bulletin* (Autumn 2006) 18-19.
13 J.M. Tangen, M.B. Thompson, D.J. McCarthy (2011) Identifying fingerprint expertise. *Psychol. Sci.* 22(8):995–997.
14 SWGFAST (2012) Document #8 Standard for the Documentation of Analysis, Comparison, Evaluation, and Verification (ACE-V) (Latent), Ver. 2.0, 11 Sept 2012. (http://www.swgfast.org/documents/documentation/121124_Standard-Documentation-ACE-V_2.0.pdf)
15 SWGFAST (2009) Standard for the Documentation of ACE-V (Latent), Ver. 1.0, 8 May 2009. (http://www.swgfast.org/documents/documentation/090508_Standard_Documentation_Latent_1.0.pdf)
16 SWGFAST (2011) Standards for the Definition and Measurement of Rates of Errors and Inappropriate Decisions in Friction Ridge Examination (Draft for Comment), Ver. 1.0, 11 Feb 2011. (http://www.swgfast.org/documents/error/110315_Rates-of-Error_DRAFT_1.0-Archived.pdf)
17 SWGFAST (2012) Document #15 Standard for the Definition and Measurement of Rates of Errors and Non-Consensus Decisions in Friction Ridge Examination (Latent/Tenprint), Ver. 2.0, 15 Nov 2012. (http://www.swgfast.org/documents/error/121124_Rates-of-Error_2.0.pdf)
18 Illinois v. Robert Morris, No. 11 CR 12889-01. Cook County Circuit Court, Illinois. 30 April 2012.