What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts

Kristy A. Martire
School of Psychology
University of New South Wales

Bethany Growns
School of Psychology
University of New South Wales

Danielle J. Navarro
School of Psychology
University of New South Wales

Abstract

Forensic handwriting examiners currently testify to the origin of questioned handwriting for legal purposes. However, forensic scientists are increasingly being encouraged to assign probabilities to their observations in the form of a likelihood ratio. This study is the first to examine whether handwriting experts are able to estimate the frequency of US handwriting features more accurately than novices. The results indicate that the absolute error for experts was lower than novices, but the size of the effect is modest, and the overall error rate even for experts is large enough as to raise questions about whether their estimates can be sufficiently trustworthy for presentation in courts. When errors are separated into effects caused by miscalibration and those caused by imprecision, we find systematic differences between individuals. Finally, we consider several ways of aggregating predictions from multiple experts, suggesting that quite substantial improvements in expert predictions are possible when a suitable aggregation method is used.

Correspondence concerning this article should be sent to A/Prof. Kristy Martire. School of Psychology, The University of New South Wales, Sydney, NSW, 2052 Australia. This work was supported by ARC DECRA Fellowship to KAM (DE140100183). Thanks to Thomas Vastrick for providing database access and thoughtful comments on an earlier draft, Jonathan Berengut for his assistance with data representation, and Brigid Betz-Stablein from Stats Central. The authors declare that they had no conflicts of interest with respect to their authorship or the publication of this article. KAM developed the study concept and obtained access to the experimental stimuli. The study was designed by KAM and BG, and programming was completed by BG. Analyses were performed by DJN, and all authors contributed to writing the manuscript.

Introduction

What makes someone an expert? On the one hand, legal scholarship and rules of evidence often cite the importance of knowledge, skill, experience, training, or education in a particular discipline (for example in Rule 702 of the *U.S. Federal Rules of Evidence*, 2016 and Section 79 of the Australian *Evidence Act (Cth)*, 1995; see Martire & Edmond, 2017). From a scientific perspective, both popular accounts (Gladwell, 2008; Ericsson & Pool, 2016) and scholarly literature on the development of expertise place an emphasis on the critical role of deliberate practice (Ericsson, Krampe, & Tesch-Römer, 1993). Nevertheless, genuine expertise is more than mere experience: it should also produce expert *performance*, characterized as "consistently superior performance on a specified set of representative tasks for a domain" (Ericsson & Lehmann, 1996, p. 277). Yet in many situations "performance" is not at all straightforward to define, as there are very often no agreed upon ground truths or gold standards that can be used as the basis for assessment (Weiss, Shanteau, & Harries, 2006). Indeed, this is the situation for many forensic science disciplines (Taroni, Aitken, & Garbolino, 2001).

An alternative approach suggests that expertise can be characterized in terms of the ability to make fine-grained discriminations in a consistent manner, as captured by measures such as the Cochran-Weiss-Shanteau (CWS) index (Weiss & Shanteau, 2003). This approach is appealing insofar as it can be applied even when objective gold standards are not available, but has one substantial drawback: it characterizes expertise in terms of the *precision* (i.e., discriminability and consistency) of expert performance rather than the *accuracy* (i.e., correctness).

One might hope that precision and accuracy are related, but in real life there are no guarantees that this is so. Whether one is considering the widespread belief in phrenology in the 19th century (Faigman, 2007), disproportionate trust in unreliable (Saks & Koehler, 2005) or unvalidated forensic methods (President's Council of Advisors on Science and Technology, 2016), beliefs in conspiracy theories (Goertzel, 1994) or "groupthink" that plagues decision making processes (Janis, 1982), it is clear that communities of people can come to considerable agreement about incorrect claims. A phrenologist might indeed produce very precise judgments, making fine-grained discriminations about a person's character in a consistent way based on their physiognomy, but that does not mean the judgments of phrenologists are sufficiently accurate to be considered *expert* in a scientific sense. From a practical standpoint, therefore, the distinction between the precision of an individual and the accuracy of their judgments is of critical importance.

In this paper we explore this question in a real world domain, using forensic handwriting expertise as our testing ground. The domain is one of considerable practical importance: forensic handwriting examinations can be used to establish the origin of a questioned sample of handwriting for legal purposes (Dyer, Found, & Rogers, 2006), and these judgments can be accorded considerable weight at trial. The task is heavily reliant on subjective judgment, with human examiners completing these assessments via visual comparison of handwriting samples (Dror & Cole, 2010). Traditionally handwriting and other feature-comparison examiners (e.g., fingerprint) have been permitted to make categorical judgments ('match' or 'no-match') without providing information about the uncertainty associated with their conclusion. Past research has suggested that forensic handwriting examiners are remarkably

accurate in their ability to make these types of authorship determinations (e.g., Sita, Found, & Rogers, 2002; Found & Rogers, 2008; Kam, Gummadidala, Fielding, & Conn, 2001).

However, in the United States, the President's Council of Advisors on Science and Technology (2016) and the National Academies of Science (2009) have both strongly criticized the traditional approach - arguing that unqualified categorical opinions are scientifically unsupportable. Consistent with these criticisms, many forensic scientists now endorse the use of likelihood ratios - the relative probability of the observations under different hypotheses as to their provenance - as the appropriate method for providing expert testimony (Aitken et al., 2011). As part of adopting this approach, handwriting examiners may be called upon to observe handwriting features and then assign probabilities to feature occurrence if, for example, two handwriting samples originated from the same versus different writers (Dror, 2016). How plausible is it that human handwriting experts are capable of producing genuinely superior performance than novices on this task?

On the one hand, there is cause for optimism: research examining visual statistical learning reveals that people can automatically and unconsciously learn statistical relationships from visual arrays given relatively limited exposure (Fiser & Aslin, 2001, 2002; Turk-Browne, Jungé, & Scholl, 2005). Further, experts have previously been found to have enhanced domain-specific statistical learning in comparison to novices (Schön & François, 2011). Indeed, it has sometimes been argued that the relevant probabilities (e.g., of seeing a particular handwriting feature given that two writing samples originated from the same versus different writers) can be estimated by the examiner based on their subjective experiences (Biedermann, Garbolino, & Taroni, 2013). On the other hand, the applied problem is essentially a probability judgment task, and there is considerable evidence that people tend not to be well calibrated at such tasks (e.g., Lichtenstein, Fischhoff, & Phillips, 1982, but see Murphy & Daan, 1984 in contexts where feedback is fast and accurate). Accordingly, there is some uncertainty as to whether forensic examiners will possess the relevant expertise in a fashion that would justify the use of such judgments in a legal context.

Our goal in this paper is to present the first empirical data examining this question, and in doing so, highlight the importance of distinguishing precision from accuracy in the assessment of expert performance. Our approach relies on a recently collected database of handwriting features (Johnson, Vastrick, Boulanger, & Schuetzner, 2016). This database was funded by the US National Institute of Justice (NIJ) to statistically estimate the frequency of handwriting features in a sample representative of the US adult population. We were able to access the estimates before they became publicly available. This allowed us to rely on a measure of "ground truth" that was not yet available to experts in the field. Using this database we were able to compare the performance of experts and novices, as well as people with high exposure to the relevant stimulus domain (US participants) and those whose experience pertains to a potentially different set of environmental probabilities (non-US participants).

Method

Participants

One-hundred and fifty participants were recruited from forensic laboratories, mailing lists and universities via email invitation. Participants not completing the experiment or

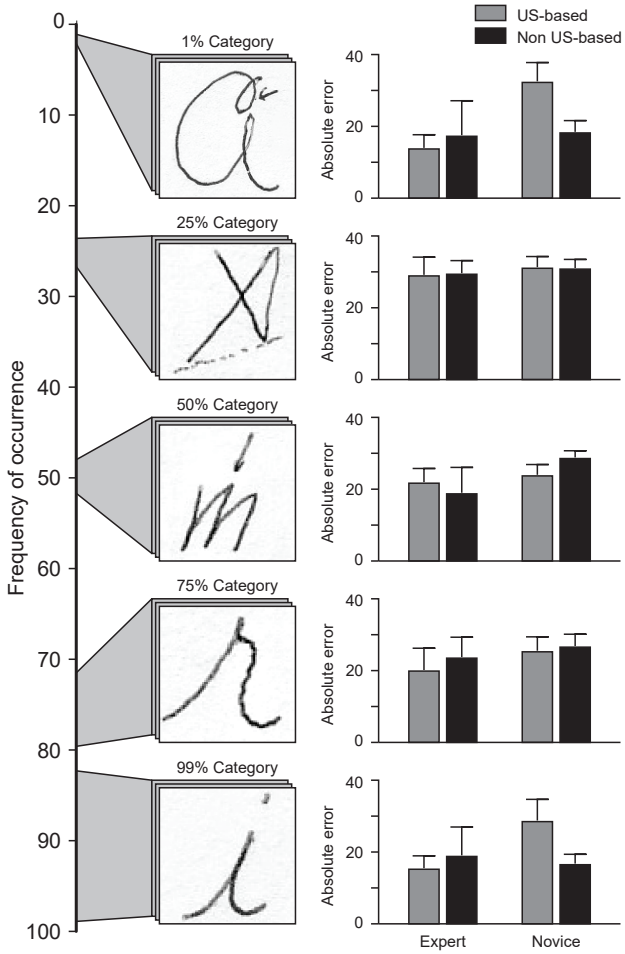

Figure 1. Stimuli and results. On each trial, participants were presented an image and verbal description of a handwriting feature exemplar. They were asked to estimate the percentage of the US population of adult writers who had the feature in their handwriting. The handwriting features varied according to case (upper or lower), style (cursive or printed) and occurrence categories (1% - top row, 25%, 50%, 75% and 99% - bottom row). The range of occurrence probabilities in each category is indicated on the left figure axis. The graphs on the right show experts' and novices' mean absolute error for each occurrence category by country (US or non-US). Error bars represent 95% within-cell confidence intervals.

not providing complete professional practice information in order for their expertise to be determined were excluded ($n = 52$). Two handwriting specialists who had not produced any expert reports or statements and one who was involved in the NIJ project were also excluded. The final sample comprised eighteen court-practicing handwriting specialists (henceforth 'experts'; $M = 149.9$ investigative and court reports from 2010-2014, Range = 9–1285; 8 US-based, 10 not) and 77 participants reporting no training, study or experience in handwriting analysis (36 US-based, 41 not). The participants that were not US-based were located in Australia (46.3%), Canada (3.2%), the Netherlands (2.1%), South Africa (1.1%) or Germany (1.1%). A $100 iTunes voucher was offered for the most accurate performance. Materials and data are available at `https://osf.io/n2g4v/`.

Stimuli

Participants were presented exemplars of handwriting features selected from the NIJ database (Johnson et al., 2016). Sixty feature exemplars (30 cursive and 30 printed handwritten forms) were selected, 12 with probabilities closest to each of five frequencies of occurrence: 1%, 25%, 50%, 75% and 99%. The range of probabilities within each category was determined by the available exemplars.

Procedure

Recruitment was necessarily time-limited and completed during the two weeks prior to the public release of the NIJ estimates. The task itself was straightforward: on every trial participants were presented a feature and asked "what percentage of the US population of adult writers have this feature in their handwriting?" using a number from 0 (never present) to 100 (present for all). On each trial, participants were shown images of handwritten letters and directed to the feature by text descriptions below (see Figure 2). After completing all 60 trials, participants provided demographic and professional information.

Results

The midpoint of the NIJ estimate range was subtracted from participant estimates to calculate absolute error for each trial. The results, averaged within-participant, are shown in Figure 3. To illustrate how these results depend on the handwriting feature itself, Figure 1 depicts the average error for the experts and novices, broken down by country of origin (US and non-US) and by rarity of feature (averaged within category).

Overall accuracy

From an applied perspective, the critical question to ask is whether the expert judges are more *accurate* than novices, and our initial (planned) analyses consider this issue first. To that end we adopted a Bayesian linear mixed model approach, using the BayesFactor package in R (Morey & Rouder, 2015), with the absolute magnitude of the error on every trial used as the dependent variable, and incorporating a random effect term to capture variability across the 95 participants and another to capture variability across the 60 features.

When analyzed in this fashion there is strong evidence (Bayes factor 39:1 against the baseline model including only the random effects) that the expert judges were more

Printed lower case ‘z’ is two strokes

What percentage of the US adult population of adult writers have this feature in their handwriting?

Please type a number between 0 and 100 in the box below.

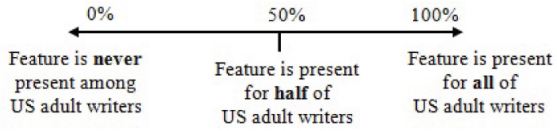

Figure 2. Screenshot illustrating the task presented to participants.

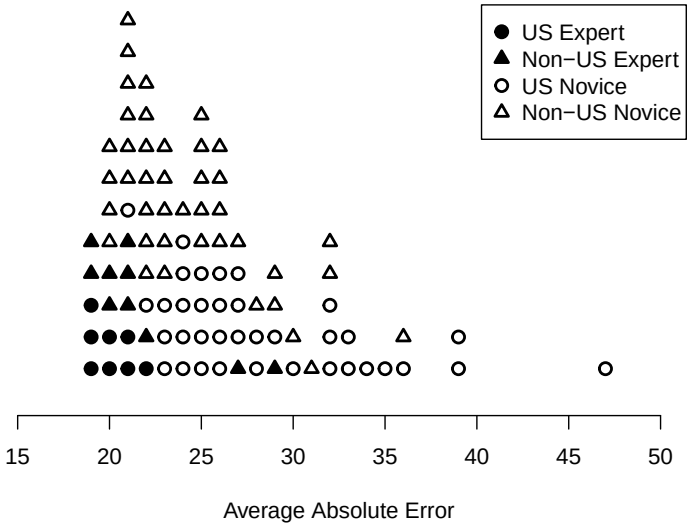

Figure 3. Histogram showing the overall accuracy of every participant, defined as the average absolute error across the 60 features. Each marker represents a single participant, with the color indicating expertise level (black = expert, white = novice) and shape indicating country (circle = US, triangle = non-US).

accurate – average error 21% on any given trial – than the novices, who produced errors of 26% on average. However, the best performing model was the 'full' model that considered all four groups (US experts, US novices, non-US experts, non-US novices) separately, with a Bayes factor of 300:1 against the baseline and 3.7:1 against a model that includes both main effects and no interaction. Consistent with this, the data show a clear ordering: the most accurate group were the US experts (20% error), followed by the non-US experts (22% error). The novices were both somewhat worse, but curiously the non-US novices performed better (24% error) than the US novices (28% error).

Estimating individual expert knowledge

While it is reassuring that handwriting experts perform better on the judgment task than novices, the specific pattern of results – in which an interaction between expertise and country is observed – requires some deeper explanation. How does it transpire that experts from the US outperformed experts from outside the US, whereas US novices appeared to perform worse than non-US novices? Do forensic handwriting examiners possess superior knowledge of the underlying probabilities, are they more precise in how they report their knowledge, or both?

To investigate these questions, we adopt a hierarchical Bayesian approach (e.g. Merkle, 2010; Lee & Danileiko, 2014) that seeks to distinguish between different kinds of errors in individual responses. As is often noted in statistics (e.g., Pearson, 1902; Cochran, 1968) the error associated with any measurement can be decomposed into different sources. For instance, researchers interested in probability judgment often estimate a "calibration curve" (e.g. Budescu & Johnson, 2011) that captures the tendency for judgments to reflect systematic bias, as opposed to the idiosyncratic variance associated with imprecise judgments. Inspired by recent work on expert aggregation models (e.g. Merkle, 2010; Lee & Danileiko, 2014), we explore these questions with the help of a hierarchical Bayesian analysis. If x_k denotes the true probability of handwriting feature k, we adopt a two-parameter version of the calibration function used by Lee and Danileiko (2014),

$$\begin{aligned} \psi_{ik} &= \delta_i \log \frac{x_k}{1-x_k} + \log \frac{c_i}{1-c_i} \\ \mu_{ik} &= \frac{\exp(\psi_{ik})}{1+\exp(\psi_{ik})} \end{aligned}$$

In this expression, μ_{ik} denotes the subjective probability that person i assigns to feature k, and it depends on two parameters: the calibration δ_i describes the extent to which all subjective probabilities for the i-th person are biased towards some criterion value, specified by the parameter c_i. This calibration curve describes the *systematic* error associated with the i-th participant. However, to capture the idea that responses y_{ik} also reflect unsystematic noise, we assume that these responses are drawn from normal distribution with mean μ_{ik}, a standard deviation σ_i that is inversely related to the *precision* $1/\sqrt{\tau_i}$ of the i-th participant, truncated to lie on the range [0,1]:

$$y_{ik} \sim \text{TruncNorm}(\mu_{ik}, \sigma_i^2, 0, 1)$$

For any given participant, the key quantities are therefore the parameters that describe their calibration function (δ_i, c_i) and the precision parameter τ_i that characterizes the amount

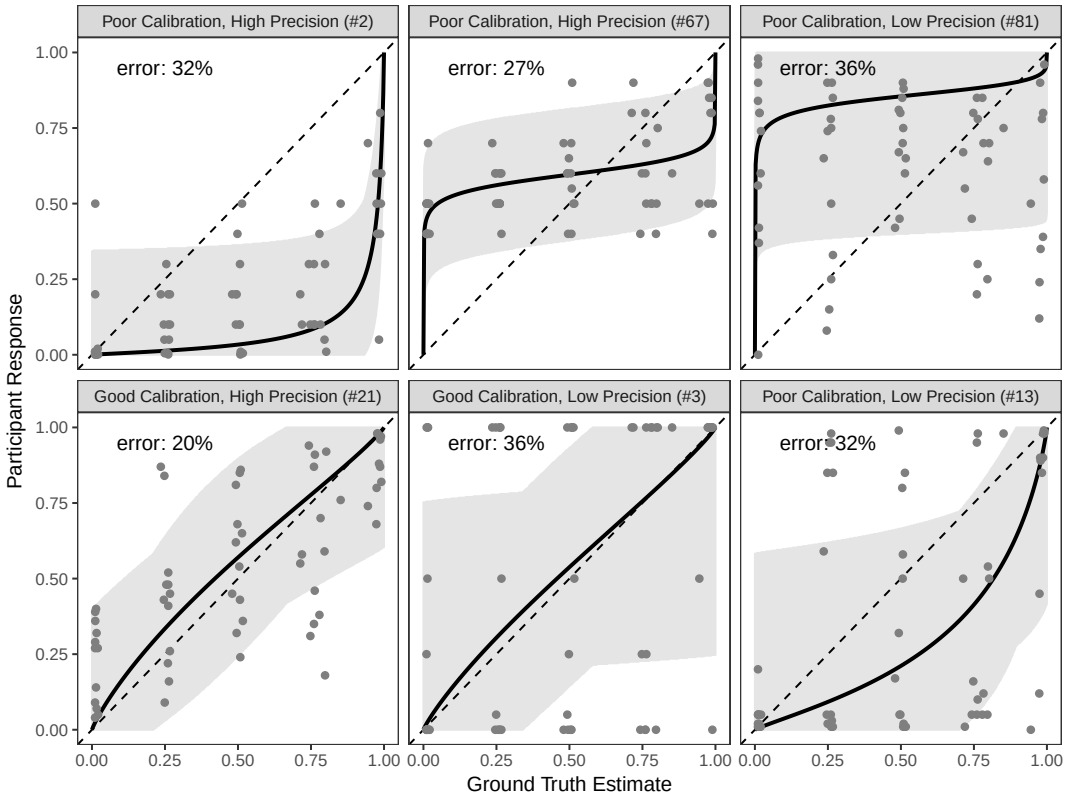

Figure 4. Illustration of how responses vary across individuals. Each panel represents data from a single person, with their overall error rate listed in the top left. Dots depict the response given by each person (y-axis) as a function of the ground truth probability for each item (x-axis). Solid lines depict the estimated calibration function that relates true probability to subjective probability for each person (dashed lines show perfect calibration), and the shaded areas depict 75% prediction intervals for each person. Some participants display good calibration (#21, #3), illustrated by the fact that the solid line lies close to the dashed line, while others do not (#2, #67, #13, #81). Some participants are relatively precise (#2, #21, #67) and have narrower prediction intervals than their less precise counterparts (#3, #13, #81). Finally, miscalibration can take different forms, with some participants responding in a way that is consistently too high (#2, #13) or too low (#81), and others responding in the middle for almost all items (#67).

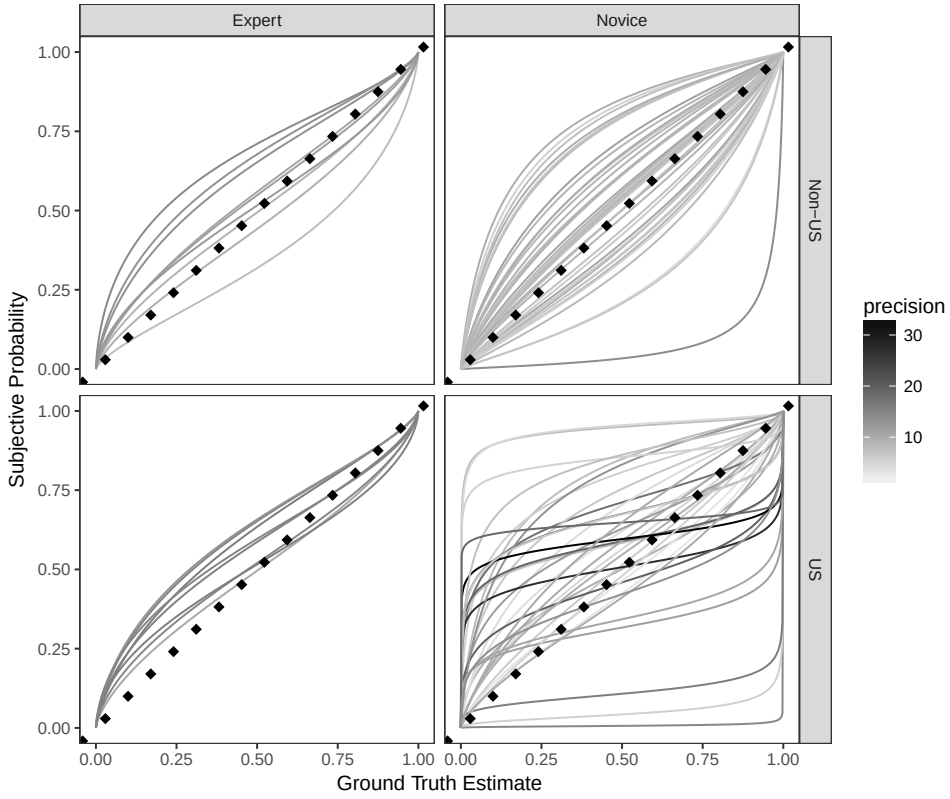

Figure 5. Estimated calibration curves for all 94 participants, plotted separately by expertise status and geographical location. The shading of each curve represents the estimated precision, with darker curves corresponding to more precise responders.

of noise in their responses. The model was implemented in JAGS (Plummer, 2003) and incorporated hierarchical priors over δ, c and τ: see Appendix for detail.

The critical characteristic of this model is that it allows us to distinguish *systematic miscalibration* from *imprecise responding*. To illustrate the importance of this distinction empirically, Figure 4 displays the raw responses, average absolute error, estimated calibration curves and degree of response precision for six participants. As is clear from inspection, both sources of error matter: participants #3 and #21 are both extremely well calibrated, but they differ considerably in their precision. Participant #3 does not make fine-grained distinctions in their responding, and as a consequence they sometimes make very large errors (e.g., rating a feature that has probability .01 as having probability 1), whereas the responses from participant #21 tend to cluster much more tightly around the true value. Similarly, while the responses of participants #2 and #67 are almost as precise as those given by participant #21, neither one is particularly well calibrated and their responses are not strongly related to the true probabilities.

Moreover, these plots highlight the manner in which the overall error rate is not always the best guide as to expertise. On the one hand, participant #21 has the lowest error rate due to the fact that they have good calibration and high precision; and conversely,

participant #81 has very high error due to their poor calibration and low precision. On the other hand, the imprecise responding of participant #3 leads to a very high error rate despite being very well calibrated. Arguably the responses of participant #3 reflect the ground truth better than participant #67 – who always provides responses in the middle of the range – even though the latter has a much lower error.

To illustrate this point more generally, Figure 5 plots the estimated calibration curve for all 94 participants, grouped by expertise and geographical location, with the shading of each curve representing the estimated precision of that participant.[1] Overall, it is clear that the expert respondents are better calibrated than novices, with the curves on the left of the figure tending to sit closer to the dotted line than those on the right. Formally, we assign each person a calibration score corresponding to the sum squared deviation (across all 60 items) between the estimated subjective probability μ_{ik} and the true value x_k, and construct 95% credible intervals for the average difference in calibration scores between the members of different groups.[2] The expert respondents were indeed better calibrated than novices, for both the US based participants (CI_{95} = [-5.4,-3.8]) and non-US participants (CI_{95} = [-1.5, -.007], though for the non-US participants the effect is modest. Similarly, we observe an difference in precision between experts and novices. Using a similar precision score that computes the sum squared deviation between the estimated subjective probability μ_{ik} and the response y_{ik} we find between experts and novices for both the US based (CI_{95} = [-2.3, -1.2]) and non-US based respondents (CI_{95} = [-2.0, -1.0]).[3]

Extracting wisdom from the forensic crowd

A natural question to ask of our data is whether there is a "wisdom of the crowd" effect (Surowiecki, 2005), in which it might be possible to aggregate the predictions of multiple participants to produce better judgments on the whole. Our sample includes a small number of real world experts and a much larger number of novices, and as Figure 5 illustrates, they differ dramatically in calibration and precision. Is it possible to aggregate the responses of these participants in a way that produces more accurate predictions than any individual expert? Does the inclusion of many poorly calibrated and imprecise novices hurt the performance of an expert aggregation model? We turn now to these questions.

Figure 6 plots the performance of two different methods of aggregating participant responses, applied to three different versions of the data: one where we included all 94 participants, one where we used responses from the 17 experts, and a third where we used

[1] In this figure, the values on x-axis are the ground truth values taken from Johnson et al. (2016). Later in the paper we introduce a version of the model that treats the feature probabilities as unknown parameters: the calibration curves produced by that model are qualitatively the same as the curves plotted here.

[2] As a subtle point, note that these intervals are constructed by treating the set of participants in each group as a fixed effect rather than a random effect, and as such we report the credible interval for the mean calibration difference for these *specific* participants, rather than attempting to draw inferences about a larger hypothetical population. This is a deliberate choice insofar as we are uncertain what larger population one ought to generalize to in this instance – it should be noted however that the credible intervals reported here necessarily correspond to a modest claim about these specific people rather than a more general claim about experts and novices.

[3] The differences between the US experts and non-US experts are more difficult to characterize: different modeling assumptions produced slightly different answers for this question, beyond the original finding that US experts performed somewhat better overall.

only the 8 US-based experts. Using the averaging method (near right), the crowd prediction is the average of each individual prediction. As the figure illustrates, this method does not yield a wisdom of crowds effect when the novices are included in the data. Using the average response of all 94 participants yields an average prediction error of 18.6%, slightly worse than the best individual participant in Figure 3 whose error was 18.5%. Once the novices are removed, the prediction error falls to 16.8% (all experts) or 16.6% (US experts only). As the left panels of Figure 6 illustrate, a decrease of 2% is about the same size as the difference between the best expert (18.5% error) and the median expert (20.4%).

One potentially problematic issue with the averaging approach is that it only works when the very poorly performing novices are removed from the data set - in many real world situations it is not known a priori who is truly expert and who is not. We consider two methods for addressing this: the first is to rely on a robust estimator of central tendency such as the median, instead of using the arithmetic mean. As shown in Figure 6, aggregation via the median produces better performance regardless of whether US experts (16.4%), all experts (16.1%) or all participants (15.7%) are included. The second method is to adopt a hierarchical Bayesian approach based on Lee and Danileiko (2014), using the more general calibration functions discussed in the previous section. The advantage of this approach is that it automatically estimates the response precision for each person, and learns in an unsupervised way which participants to weight most highly. As illustrated in the far right of Figure 6, this model makes better predictions than the averaging method or the median response method: using only the US experts the prediction error is 15.1%, which falls slightly to 14.9% when all experts are used, and improves slightly further to 14.7% when the novices are included. In other words, not only is the hierarchical Bayes method robust to the presence of novices, it is able to use their predictions to perform better than it does if only the experts are considered and yields better estimates than simple robust methods such as median response.

Finally, to understand why the Bayesian aggregation model performs better than the simpler averaging method, Figure 7 plots the model based estimates against the average human response (left) and compares both of these prediction methods to the ground truth estimated from the NIJ data (middle and right panels). As shown in the left panel, the main thing that the model has learned is to transform the average estimates via a non-linear calibration function that closely resembles curves used in standard theoretical models (Prelec, 1998), though as illustrated in Figure 4 individual subject calibration curves often depart quite substantially from this shape. This has differential effects depending on the true frequency of the handwriting feature in question: for very rare and very common features, the model reverses the regression to the mean effect, and so the Bayesian model produces much better estimates for these items (see also Satopää et al., 2014). The overall result is that the model estimates produce a much better prediction about the ground truth (middle panel) than the averaging method (right panel).

Conclusions

The assessment of expert performance is a problem of considerable importance. Besides the obvious theoretical questions pertaining to the nature of expertise (e.g. Weiss & Shanteau, 2003; Ericsson & Lehmann, 1996), a variety of legal, professional and industrial fields are substantially reliant on human expert judgment. For the specific domain of foren-

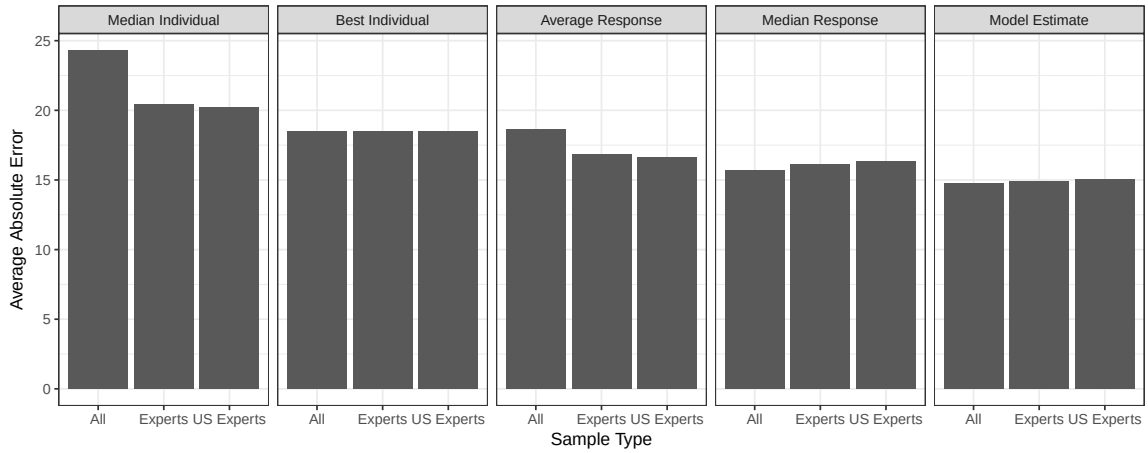

Figure 6. Performance of the aggregation model (far right), when compared to simpler aggregation approaches that averages responses (middle) or takes the median response (near right). For comparison purposes, both are plotted against the performance of the best individual participant (near left), and the median performance of all respondents (far left). Within each panel, three versions are plotted: one where we included all 94 participants, one where we used responses from the 17 experts, and a third where we used only the 8 US-based experts.

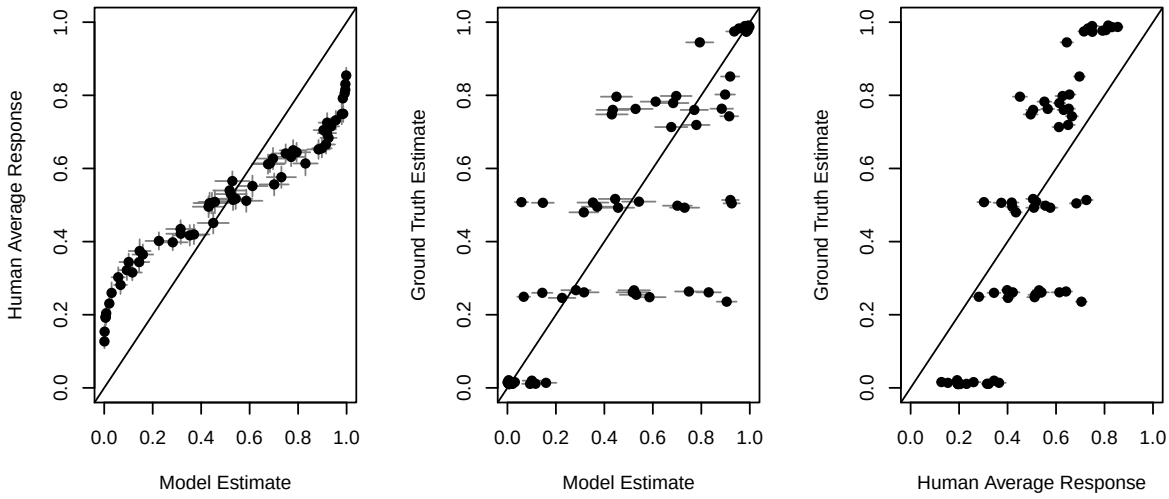

Figure 7. Aggregating expert predictions. Pairwise comparisons between the ground truth estimated from the NIJ data, the average response of human participants, and the estimates extracted via the hierarchical Bayesian model.

sic handwriting examination the legal implications are the most pressing. In this respect our findings are mixed.

On the one hand, there is some evidence that handwriting experts will be able to estimate the frequency of occurrence for handwriting features better than novices. However, even the single best performing participant produced an average deviation of 18.5% from the true value. On the other hand, this number is considerably lower than would be expected by chance (25%) if people possessed no relevant knowledge and simply responded with .5 on every trial, and using modern Bayesian methods to aggregate the predictions of the experts this can be reduced further to 14.7% error.

In short, these results provide some of the first evidence of naturalistic visual statistical learning in the context of forensic feature comparison. However, we suggest that a cautious approach should be taken before endorsing the use of experience-based likelihood ratios for forensic purposes in the future.

References

Aitken, C., Berger, C. E. H., Buckleton, J. S., Champod, C., Curran, J., Dawid, A., & Jackson, G. (2011). Expressing evaluative opinions: A position statement. *Science and Justice*, *51*, 1-2.

Biedermann, A., Garbolino, P., & Taroni, F. (2013). The subjectivist interpretation of probability and the problem of individualisation in forensic science. *Science and Justice*, *53*, 192-200.

Budescu, D. V., & Johnson, T. R. (2011). A model-based approach for the analysis of the calibration of probability judgments. *Judgment and Decision Making*, *6*, 857–869.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics*, *10*(4), 637–666.

Dror, I. E. (2016). A hierarchy of expert performance. *Journal of Applied Research in Memory and Cognition*, *5*(2), 121–127.

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

Dyer, A. G., Found, B., & Rogers, D. (2006). Visual attention and expertise for forensic signature analysis. *Journal of Forensic Sciences*, *51*(6), 1397–1404.

Edwards, H., & Gotsonis, C. (2009). *Strengthening forensic science in the United States: A path forward*. Washington, DC: National Academies Press.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual review of psychology*, *47*(1), 273–305.

Ericsson, K. A., & Pool, R. (2016). *Peak: Secrets from the new science of expertise*. Houghton Mifflin Harcourt.

Faigman, D. L. (2007). Anecdotal forensics, phrenology, and other abject lessons from the history of science. *Hastings Law Journal*, *59*, 979-1000.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*, 499-503.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458-467.

Found, B., & Rogers, D. (2008). The probative character of forensic handwriting examiners' identification and elimination opinions on questioned signatures. *Forensic Science International*, *178*(1), 54–60.

Gladwell, M. (2008). *Outliers: The story of success*. Hachette UK.

Goertzel, T. (1994). Belief in conspiracy theories. *Political Psychology*, 731–742.

Janis, I. L. (1982). *Groupthink: Psychological studies of policy decisions and fiascoes* (Vol. 349). Houghton Mifflin Boston.

Johnson, M. E., Vastrick, T. W., Boulanger, M., & Schuetzner, E. (2016). Measuring the frequency occurrence of handwriting and handprinting characteristics. *Journal of Forensic Sciences*.

Kam, M., Gummadidala, K., Fielding, G., & Conn, R. (2001). Signature authentication by forensic document examiners. *Journal of Forensic Science*, *46*(4), 884–888.

Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, *9*(3), 259-273.

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Martire, K. A., & Edmond, G. (2017). Rethinking expert opinion evidence. *Melbourne University Law Review*, *40*, 967-998.

Merkle, E. C. (2010). Calibrating subjective probabilities using hierarchical bayesian models. In *International conference on social computing, behavioral modeling, and prediction* (pp. 13–22).

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.12-2)

Murphy, A. H., & Daan, H. (1984). Impacts of feedback and experience on the quality of subjective probability forecasts. comparison of results from the first and second years of the zierikzee experiment. *Monthly Weather Review*, *112*(3), 413–423.

Pearson, K. (1902). On the mathematical theory of errors of judgment, with special reference to the personal equation. *Philosophical Transactions of the Royal Society of London, Series A*, *198*, 235–299.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 125).

Prelec, D. (1998). The probability weighting function. *Econometrica*, *66*, 497–527.

President's Council of Advisors on Science and Technology. (2016). *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods.* Washington, DC: Executive Office of the President of the United States.

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*(5736), 892–895.

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H.

(2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, *30*(2), 344–356.

Schön, D., & François, C. (2011). Musical expertise and statistical learning of musical and linguistic structures. *Frontiers in Psychology*, *2*(167), 1-9.

Sita, J., Found, B., & Rogers, D. K. (2002). Forensic handwriting examiners' expertise for signature comparison. *Journal of Forensic Science*, *47*(5), 1–8.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Taroni, F., Aitken, C., & Garbolino, P. (2001). De finetti's subjectivism, the assessment of probabilities and the evaluation of evidence: a commentary for forensic scientists. *Science & Justice*, *41*(3), 145–150.

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, *134*(4), 552-564.

Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors*, *45*(1), 104–116.

Weiss, D. J., Shanteau, J., & Harries, P. (2006). People who judge people. *Journal of Behavioral Decision Making*, *19*(5), 441–454.

Appendix

The hierarchical Bayesian calibration model used in the main text is specified as follows. Let $x_k \in [0,1]$ denote the true frequency of the k-th feature, and let $y_{ik} \in [0,1]$ denote the judged frequency of feature k produced by i-th participant. The main text describes a two-parameter calibration function,

$$\begin{aligned} \psi_{ik} &= \delta_i \log \frac{x_k}{1-x_k} + \log \frac{c_i}{1-c_i} \\ \mu_{ik} &= \frac{\exp(\psi_{ik})}{1+\exp(\psi_{ik})} \end{aligned}$$

and we assume that the response is sampled from a truncated normal distribution

$$y_{ik} \sim \text{TruncNorm}(\mu_{ik}, \sigma_i^2, 0, 1)$$

where the standard deviation σ_{ik} is related to the precision via $\tau_{ik} = 1/\sigma_{ik}^2$, and notation above indicates that the distribution is truncated below at 0 and above at 1.

To accommodate individual and group differences we assume that the precision τ_i, calibration δ_i and criterion c_i parameters are sampled from a population distribution that may have slightly parameters as a function of the group g_i (e.g., US expert) to which the expert belongs. For each group g the calibration parameter δ is sampled from a Gaussian distribution with unknown mean $\mu_{\delta,g}$ and precision $\tau_{\delta,g}$, truncated to lie between 0 and 1. A similar approach is applied to the precision and criterion parameters, yielding the following model at the group level:

$$\begin{aligned} \delta_i | g_i = g &\sim \text{TruncNorm}(\mu_{\delta,g}, 1/\tau_{\delta,g}, 0, 1) \\ c_i | g_i = g &\sim \text{TruncNorm}(\mu_{c,g}, 1/\tau_{c,g}, 0, 1) \\ \tau_i | g_i = g &\sim \text{TruncNorm}(\mu_{\tau,g}, 1/\tau_{\tau,g}, 0, \infty) \end{aligned}$$

To capture the intuition that the various groups may be somewhat similar to one another, the group level parameters are assumed to be drawn from a higher level population distribution, again assumed to be a truncated normal distribution

$$\begin{aligned}
\mu_{\delta,g} &\sim \text{TruncNorm}(\mu'_1, 1/\tau'_1, 0, 1) \\
\mu_{c,g} &\sim \text{TruncNorm}(\mu'_2, 1/\tau'_2, 0, 1) \\
\mu_{\tau,g} &\sim \text{TruncNorm}(\mu'_3, 1/\tau'_3, 0, \infty) \\
\tau_{\delta,g} &\sim \text{TruncNorm}(\mu'_4, 1/\tau'_4, 0, \infty) \\
\tau_{c,g} &\sim \text{TruncNorm}(\mu'_5, 1/\tau'_5, 0, \infty) \\
\tau_{\tau,g} &\sim \text{TruncNorm}(\mu'_6, 1/\tau'_6, 0, \infty)
\end{aligned}$$

Finally, non-informative hyperpriors were adopted to describe the prior over the population level parameters μ' and τ': Gamma(.01,.01) distributions for parameters lower bounded at zero, and uniform priors for parameters that lie between 0 and 1.