Psychometrics for Forensic Fingerprint Comparisons

Amanda Luby[1], Anjali Mazumder[2], and Brian Junker[3]

[1] Swarthmore College, Swarthmore, PA, USA
aluby1@swarthmore.edu,
www.swarthmore.edu/NatSci/aluby1
[2] The Alan Turing Institute, London, UK
[3] Carnegie Mellon University, Pittsburgh, PA, USA

Abstract. Forensic science often involves the evaluation of crime-scene evidence to determine whether it matches a known-source sample, such as whether a fingerprint or DNA was left by a suspect or if a bullet was fired from a specific firearm. Even as forensic measurement and analysis tools become increasingly automated and objective, final source decisions are often left to individual examiners' interpretation of the evidence. Furthermore, forensic analyses often consist of a series of steps. While some of these steps may be straightforward and relatively objective, substantial variation may exist in more subjective decisions. The current approach to characterizing uncertainty in forensic decision-making has largely centered around conducting error rate studies (in which examiners evaluate a set of items consisting of known-source comparisons) and calculating error rates aggregated across examiners and identification tasks. We propose a new approach using Item Response Theory (IRT) and IRT-like models to account for differences in examiner behavior and for varying difficulty among identification tasks. There are, however, substantial differences between forensic decision-making and traditional IRT applications such as educational testing. For example, the structure of the response process must be considered, answer keys for comparison tasks do not exist, and information about participants and items is not available due to privacy constraints. In this paper, we provide an overview of forensic decision-making, outline challenges in applying IRT in practice, and survey some recent advances in the application of Bayesian psychometric models to fingerprint examiner behavior.

Keywords: Item Response Theory, Forensic Science, Bayesian Statistics

1 Introduction

"Forensic science" is a broad field that consists of many different scientific disciplines that are used in a legal context. These scientific disciplines range from highly objective, such as single-source DNA analysis, to highly subjective, such as bite mark analysis. Forensic science relies on forensic examiners, who are responsible for determining whether a piece of evidence left at a crime scene came from a particular source. Depending on the type of evidence, this process may be nearly automatic and consistent across examiners or vary considerably depending on the examiner performing the analysis. For many disciplines, examiners report their results as an expert opinion of one of three outcomes: the suspect is the source of the evidence (known as an identification or individualization)[4], the suspect is not the source of the evidence (known as an exclusion), or that the analysis was inconclusive (Stern, 2017).

This work focuses on fingerprint evidence, in which a forensic examiner compares a *latent* fingerprint (e.g. from a crime scene) to one or more *reference* prints to determine whether they came from the same source or not. The standard operating procedure for analyzing fingerprint evidence is a series of steps known as ACE-V (Analysis, Comparison, Evaluation, Verification), but each step in the ACE-V process is a complex task involving many different factors (see, e.g., OSAC, 2019 for details), and forensic examiners may vary in their approach to the ACE-V process. They may have different standards for the quality or clarity of latent fingerprint needed to perform an analysis, may select different fingerprint features (called minutiae) on which to base a comparison, and may have different thresholds for the degree of similarity required to declare an individualization (or exclusion).

The current approach to characterizing uncertainty in examiner decisions has focused on the calculation of aggregated error rates across all examiners and identification tasks. This approach is not ideal for comparing examiner performance, as examiner decisions are not always unanimous, and error rates are likely to vary across identification tasks depending on the difficulty of the comparison. The variation in examiner decisions alongside the variation in task difficulty makes this application conducive for Item Response Theory (IRT) and related psychometric models (Kerkhoff et al, 2015; Luby et al, 2020).

However, standard IRT approaches must be adapted for this type of data. First, responses are not keyed as 'correct' or 'incorrect' by the test provider. While we may infer that an individualization of a same-source print is 'correct', and an exclusion of a same-source print is 'incorrect' (and vice-versa for different-source prints), it is unclear how 'inconclusive' responses should be treated. For example, an inconclusive on a low-quality print may be considered the 'correct' decision, but an inconclusive on a high-quality print may be considered an 'incorrect' decision since a potential individualization or exclusion was missed. Second, fingerprint comparisons consist of a series of sequential steps. Collapsing the decisions made at each of these steps into a single response ignores the conditional structure of the responses and results in the loss of information about variation at each step of the process. We propose the use of the Item Response Trees framework (IRTrees, De Boeck and Partchev 2012) as a solution to these issues.

[4] An 'individualization' is an 'identification' to the global exclusion of all others (OSAC, 2017). Following Ulery et al (2011), we use 'individualization' and do not distinguish between 'individualization' and 'identification'.

The remainder of the paper is organized as follows. In Section 2, we introduce the FBI Black Box Study (Ulery et al, 2011), which is the source of the data used throughout the paper. In Section 3, we introduce the IRTrees framework and a model for the fingerprint comparison task. Results are briefly described in Section 4, and limitations and future work are discussed in Section 5.

2 Data

The FBI Black Box study (Ulery et al, 2011) was the first large-scale study performed to assess the accuracy and reliability of fingerprint examiners' decisions in the United States. 169 latent print examiners were recruited for the study, and each participant was assigned roughly 100 items from a pool of 744. Each item consisted of a *latent print* (fingerprint of unknown source lifted from, e.g., a crime scene) and a *reference print* (fingerprint of known source taken under idealized conditions). The latent prints were designed to include a range of features and quality similar to those seen in casework and to be representative of searches from an automated fingerprint identification system.

The study provided an estimate of the aggregated false positive rate (0.1%) and false negative rate (7.5%) in casework. In addition, each recorded response to an item consists of results from the following decisions:

1. Latent evaluation: the examiner's evaluation of whether the crime scene print is of *No Value*, *Value for Exclusion Only*, or *Value for Individualization.*
2. Source decision: the examiner's decision of whether the pair of prints is an *Exclusion* (different sources), *Individualization* (same source), or *Inconclusive.*
3. If inconclusive, one of:
   – *Close*: The correspondence of features is supportive of the conclusion that the two impressions originated from the same source, but not to the extent sufficient for individualization
   – *Insufficient Information*: Potentially corresponding areas are present, but there is insufficient information present.
   – *No Overlap*: No overlapping areas between the latent and reference print
4. If exclusion, one of:
   – *Pattern*: The exclusion determination could be made on fingerprint pattern class (the overall shape of the fingerprint ridges) and did not require the use of minutiae (the small details in the fingerprint).
   – *Minutiae*: The exclusion determination required the use of minutiae .
5. Difficulty (Five-point scale)

Note that due to conflicting responses in the latent evaluation stage, we do not distinguish between *value for individualization* and *value for exclusion only* for this analysis, and treat the latent evaluation as a binary response (*Has value* vs *No value*) instead. We also base our analysis on OSAC, 2019, and pool the *Insufficient Information* and *No Overlap* inconclusives into one category.

While the study emphasized estimating casework error rates and therefore focused on the source decision, important trends in examiner behavior are also present in the other decisions. For example, latent print examiners vary in their tendencies towards ‘no-value’ and ‘inconclusive decisions’. Figure 1 shows the distribution of the number of inconclusive and no value decisions reported by each examiner. Although most examiners report between 20-40 inconclusives and 15-35 ‘no value’ responses, some examiners report as much as 60 or as few as 5.

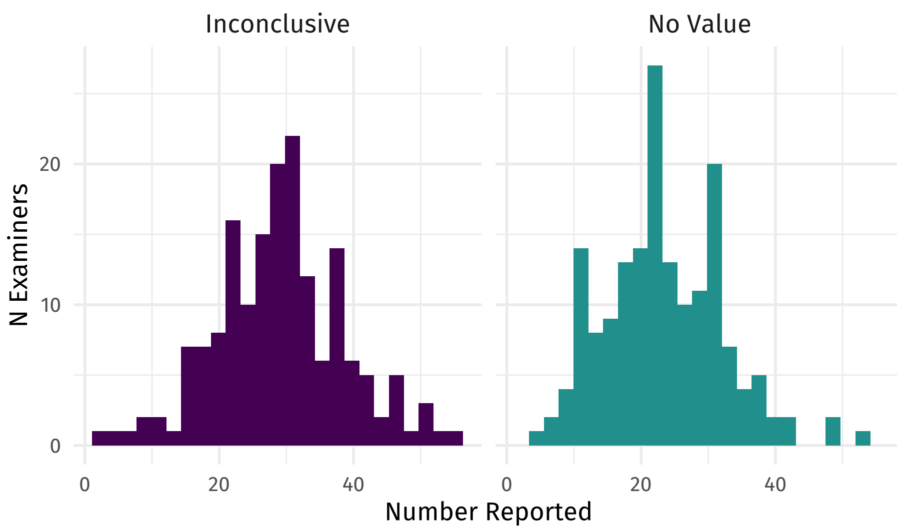


Fig. 1. Number of inconclusive and no value decisions reported by each examiner.

Furthermore, there are some items which examiners largely agree on, and other items where there is substantial disagreement. Figure 2 shows an example of one high-disagreement item (left) and one low-disagreement item (right). Each column represents one of the sub-decisions made for each item assessment: (1) latent evaluation, (2) source decision, and (3) reason for the decision. We note that, even for the item on the right for which examiners largely agreed, there is still some disagreement in both the source decision and the reason.

By modeling these responses explicitly, we can assess individual differences among examiners in their tendencies to make latent evaluations, source decisions, and reasons for decisions. Similarly, we can measure the variation among items for each stage in the decision-making process.

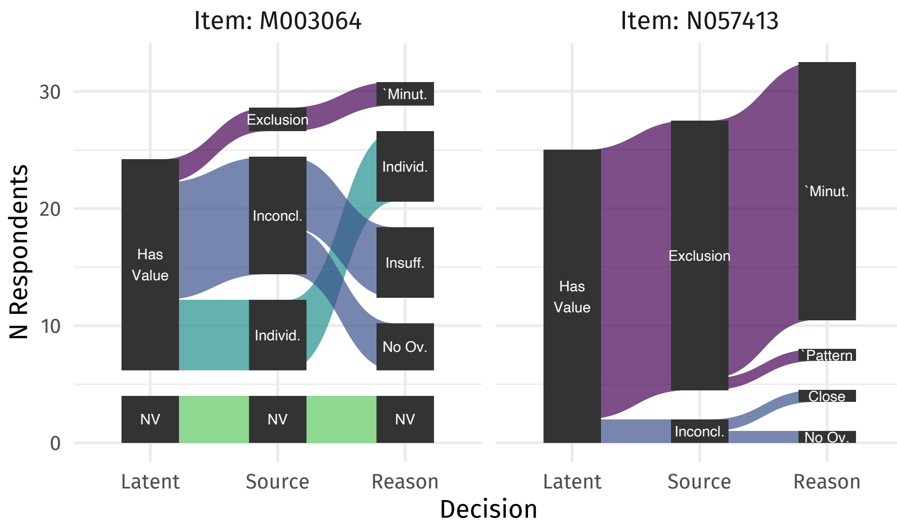
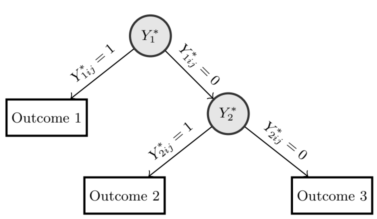


Fig. 2. An illustration of how examiners responded to a high-disagreement item (left) and low-disagreement item (right) for each of three sub-decisions (latent value, source decision, reason for decision).

3 Item Response Trees

Item Response Trees (IRTrees, De Boeck and Partchev 2012) use decision trees to describe hypothesized cognitive processes, where the leaves are the final observed outcome. The IRTree formulation can represent a wide variety of response formats and response processes, easily adapted for binary responses, unipolar scales, bipolar scales, and Likert responses (Jeon and De Boeck, 2016). In the forensic science setting, IRTrees are useful for representing the sequential decision-making process explicitly (Luby et al, 2020).



$$
\begin{aligned}
P(Y_{ij} = \text{Outcome } 1) &= P(Y_{1ij}^* = 1) \\
P(Y_{ij} = \text{Outcome } 2) &= P(Y_{1ij}^* = 0) \\
&\quad \times P(Y_{2ij}^* = 1 | Y_{1ij}^* = 0) \\
P(Y_{ij} = \text{Outcome } 3) &= P(Y_{1ij}^* = 0) \\
&\quad \times P(Y_{2ij}^* = 0 | Y_{1ij}^* = 0)
\end{aligned}
$$

Fig. 3. Example IRTree model tree structure (left) and outcome probabilities (right).

Figure 3 illustrates a basic IRTree model for a response with three possible outcome categories (e.g. $Y = 1, 2, 3$), where Y_1^* and Y_2^* are nodes constructed to represent internal decisions that lead to each of three outcomes. If Y_{ij} denotes the response of participant i ($i = 1, ..., I$) to item j ($j = 1, ..., J$), Y^*_{1ij} denotes the choice of left or right branch at Y^*_1 for person i at item j.

The probability of choosing the left branch at node 1 (Y^*_1 in Figure 3) can be modeled using standard IRT models. We use the Rasch model at binary nodes for interpretability and computational convenience: $P(Y^*_{kij} = 1) = \text{logit}^{-1}(\theta_{ki} - b_{kj})$, where θ_{ki} denotes the latent trait involved with choosing the left branch at node k for person i and b_{kj} is the corresponding Rasch parameter for item j. In a standard Rasch model for correct/incorrect outcomes, the item parameters (b_j) correspond to the difficulty of the item, where higher values of b_j decrease the probability that $Y_{ij} = 1$. When IRTree branch decisions do not correspond to incorrect/correct choices, the item parameters represent an "item tendency" towards one branch over the other rather than difficulty.

The model for the probability of choosing the left branch at Y^*_2 is similar, except it is conditional on Y^*_1 being equal to zero (i.e. we model $P(Y^*_{2ij} = 1|Y^*_{1ij} = 0)$ instead of $P(Y_{2ij} = 1)$). The probability of each observed response (Outcome 1, 2, or 3) is then the product of the probabilities of the internal branches leading to each leaf in the tree (Figure 3 right).

3.1 Model for Fingerprint Comparisons

We use an IRTree model constructed using the OSAC Process Map (OSAC, 2019). The constructed IRTree is a necessary simplification of the process map based on the available data in the Black Box study. There are many decisions represented in the Process Map, but there is no way to reconstruct many decisions based on the responses that were recorded in the Black Box study. We also note that not every examiner uses the Process Map for every decision, and that processes may vary by agency.

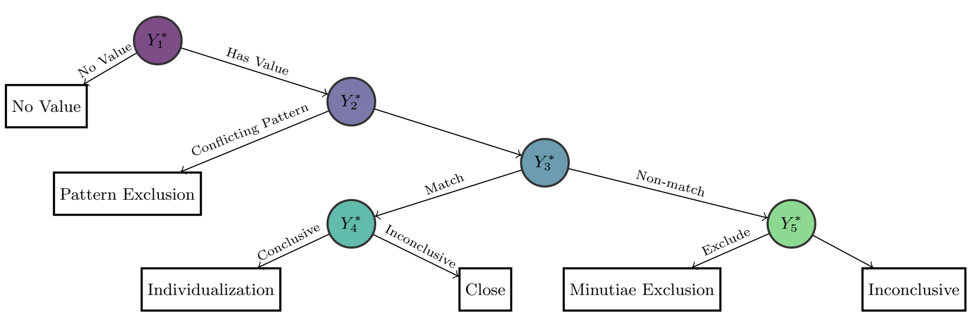


Fig. 4. The *OSAC Process Map* IRTree. Nodes are colored to match the corresponding plots in Figure 5.

Each node is parameterized using a Rasch model with $b_{kj} = \beta_{0k} + \beta_{1k} X_j + \epsilon_{kj}$, where $X_j = 1$ if item j is a true same-source pair and 0 if item j is a different-source pair.

We take a Bayesian approach to estimation, which allows us to estimate posterior distributions for all participant and item parameters simultaneously. The IRTree model was implemented in Stan (Stan Development Team, 2018b,a) using R (R Core Team, 2013). Multivariate normal distributions were chosen for $\boldsymbol{\theta}$ and $\mathbf{b}$. Other parameter distributions were chosen based on recommended priors for efficiency, and all code is publicly available[5].

$$
\begin{aligned}
\boldsymbol{\theta}_i &\sim MVN_5(\mathbf{0}, \boldsymbol{\sigma_\theta} L_\theta L_\theta' \boldsymbol{\sigma_\theta}),\\
L_\theta &\sim LKJ(4),\\
\boldsymbol{\sigma_b} &\sim \text{Half-Cauchy}(0, 2.5),\\
\boldsymbol{b}_j &\sim MVN_5(\boldsymbol{\beta}\mathcal{X}_j, \boldsymbol{\sigma_b} L_b L_b' \boldsymbol{\sigma_b}),\\
L_b &\sim LKJ(4),\\
\boldsymbol{\sigma_b} &\sim \text{Half-Cauchy}(0, 2.5),\\
\beta_k &\sim N(0, 5).
\end{aligned}
$$

Here $\mathcal{X}_j$ is the column vector $(1, X_j)'$, $\boldsymbol{\beta} = (\boldsymbol{\beta_1}, ..., \boldsymbol{\beta_5})$ is the 5×2 matrix whose k^{th} row is (β_{0k}, β_{1k}), and $\boldsymbol{\sigma_b}$ is a 5×5 diagonal matrix with $\sigma_{1b}, ..., \sigma_{5b}$ as the diagonal entries; $\boldsymbol{\sigma_\theta}$ in the previous line is defined similarly. The Stan modeling language does not rely on conjugacy, so the Cholesky factorizations (L_θ and L_b) are modeled instead of the covariance matrices for computational efficiency.

4 Results

The IRTree model introduced in Section 3.1 is complex and results in 5 parameters per person and per item, in addition to hyperparameters. We focus on two aspects of the results of the model: (1) the magnitude and uncertainty of participant parameters and (2) using item parameters to generate an "answer key".

4.1 Participant Parameters

For each of the 169 participants (indexed by i), the IRTree model estimates five parameters: a 'no value' tendency (θ_{1i}), a 'pattern exclusion' tendency (θ_{2i}), a 'match' tendency (θ_{3i}), an 'individualization' tendency (θ_{4i}), and a 'minutiae exclusion' tendency (θ_{5i}). Each of these θ estimates correlates with an observed outcome (e.g. θ_{1i} correlates with percent of *No Value* decisions) but also accounts for the corresponding item tendencies. For example, θ_1 represents the tendency of an examiner to choose *no value* after accounting for the subset of items that

[5] github.com/aluby/imps2020

they were shown. Figure 5 shows the five θ estimates for each examiner (with 95% posterior intervals) as compared to each examiner's proficiency estimate from a Rasch model fitted to scored data.

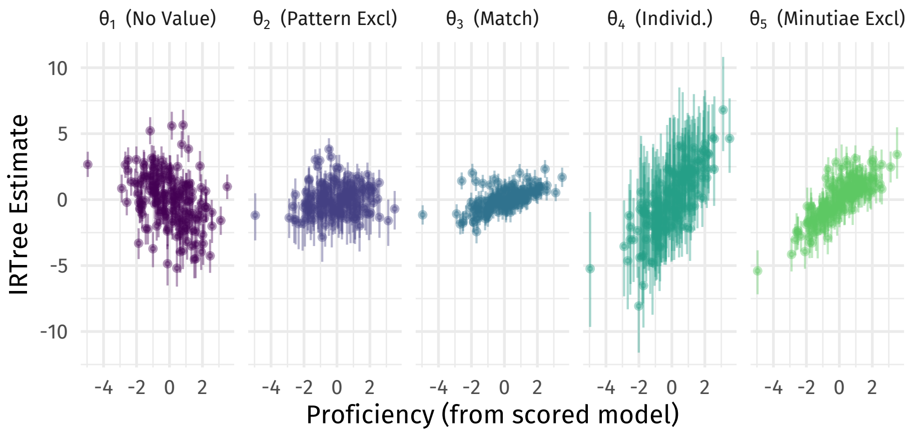


Fig. 5. Number of inconclusive and no value decisions reported by each examiner.

First, we note that estimated proficiency under a Rasch model is not sufficient for understanding examiner behavior. As outlined in Section 1 and 2, fingerprint comparisons are a complex task consisting of a series of steps. Any mapping from the original responses to a binary response necessarily results in the loss of information. Furthermore, there is no designated 'answer key' for the Black Box items, and it is unclear how 'inconclusive' or 'no value' responses should be treated.

Figure 5 demonstrates that even though the IRTree does not require any responses to be scored as correct or incorrect, θ_4 and θ_5 (and to some extent the other parameters) are still correlated with proficiency from a Rasch model. That is, we can still identify examiners who often correctly individualize or exclude (corresponding to more positive θ_4 and θ_5 estimates), as well as those who make more false individualizations and exclusions (corresponding to more negative θ_4 and θ_5 estimates).

Furthermore, the 'match' tendency estimates (θ_3) are the least extreme in magnitude of all of the θ estimates. Examiners are therefore unlikely to disagree on whether a pair of fingerprints is more likely a 'match' or a 'non-match'(the Y_3^* split of the IRTree model in Figure 4), but *do* disagree on the level of certainty in such a decision (i.e. individualization vs close, Y_4 or minutiae exclusion vs inconclusive, Y_5). This is consistent with previous work that examiners 'willingness to respond' drives much of the disagreement (Dror and Langenburg, 2019).

Finally, we note that there is substantial variation in the 'no value' tendency θ_1. We observe a slight negative correlation between θ_1 and proficiency from a scored model, and that examiners with negative θ_1 estimates (less likely to rate items as *no value*), tend to have positive θ_4 and θ_5 estimates (more likely to rate items as *individualization* or *minutiae exclusion*), providing a link between some of the ‘willingness to respond’ parameters. The posterior intervals for θ_1 are also noticeably smaller than, e.g., individualization tendency (θ_4), likely due to more observations at earlier nodes in the IRTree.

4.2 Generating an ‘answer key’ from item parameters

Using the parameter estimates from each item, we can also estimate the probability of observing each response for a hypothetical “unbiased” examiner. For example, a completely unbiased examiner would have $\theta_k = 0$ for all k nodes in the IRTree, resulting in responses that are totally driven by the item parameters. If such an examiner responded to all 744 items in the FBI “Black Box” study, we can calculate the predicted response to each item. These results could be used as an “answer key” to identify potentially problematic responses since correct and incorrect responses are not keyed by the FBI.

Table 1 compares the IRTree answer key described above to a modal answer key, where the expected answer to each item is determined by the most popular response to that item. We see that the answer keys largely agree, with both keys labeling very few items as *Pattern Exclusion* or *Close* inconclusive. The most disagreement between the answer keys occurs when the modal answer key predicted a *No Value* or an *Other Inconclusive* and the IRTree model obtained a different label (the first and fourth column, respectively), and when the IRTree model labeled an item with an *Other Inconclusive* (the fourth row).

Table 1. A comparison of how each item was keyed by the IRTree answer key (rows) and the Modal answer key (columns). While the two answer keys agree on most items (corresponding the diagonal entries), there is some disagreement when the modal answer key labeled items as No Value (first column) or inconclusive (fourth column).

		Modal Answer Key					
		No Value	Individualization	Close	Inconclusive	Minutiae Exclusion	Pattern Exclusion
IRTree Answer Key	No Value	175			2		
	Individualization	1	173		2		
	Close			24	1		
	Inconclusive	13		7	170		
	Minutiae Exclusion	1			4	154	2
	Pattern Exclusion	2					13

While Table 1 compares the two answer keys across items, we can also compare the results of the answer keys across participants. Figure 6 shows the observed score (% Correct) for each participant under the IRTree Answer Key and

Modal Answer Key. If there was perfect correspondence across the two answer keys, all points would be located on the dashed diagonal line. While some participants receive slightly higher or lower scores under the IRTree answer key than they do under the modal answer key, the scores do not change substantially or in a systematic way. For this setting, we prefer the IRTree framework since the expected responses account for patterns in examiner behavior. For a further discussion of IRTree-generated answer keys and their relationship with other methods such as cultural consensus theory, see Luby (2019).

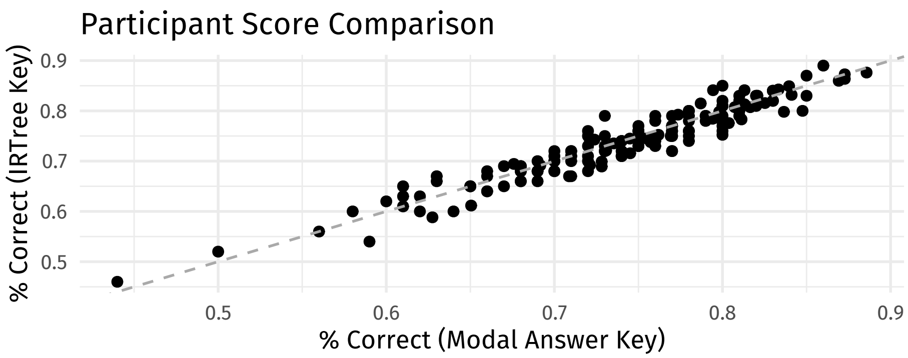


Fig. 6. Each participant's observed score under the IRTree answer key compared to their observed score under the modal answer key. Perfect correspondence is indicated by the dashed line.

5 Discussion

The current approach to characterizing uncertainty in forensic decision-making is largely focused on estimating aggregated error rates across examiners and identification tasks. We have proposed a new approach using IRTrees to account for differences in examiner behavior at different points in their decision-making process, and how estimated parameters can be used to generate an answer key to identify potentially problematic responses. Although there are many items with substantial variation in the responses, most items were found to have a clear expected answer. Examiners should receive feedback not only when they make a false identification or exclusion, but also when mistaken 'inconclusive' or 'no value' decisions are made. In order to provide such feedback, expected answers must first be generated.

There are, however, limitations to the types of analyses we can perform with this data, particularly in explanatory modeling. For example, rich survey data was collected alongside responses in the Black Box study (e.g. type of training, years of experience, etc.) but survey responses were not linked to test responses to maintain the confidentiality of participants. Furthermore, unlike traditional IRT applications, there are also privacy concerns regarding the items themselves. Each item consists of a pair of images of fingerprints, which by nature are identifiable and cannot be publicly released. This complicates explanatory modeling for participants *and* for items.

Following the Black Box study, there was a series of follow-up studies performed using the same set of participants, and we plan to expand our analyses to include these results. The first was a 'repeatability' study (Ulery et al, 2012), in which 72 participants of the original Black Box study were asked to re-analyze 25 questions seven months after the original study, which provides a unique opportunity to validate conclusions on truly out-of-sample data. The 'White Box' study (Ulery et al, 2014) asked examiners to annotate features, image clarity, and correspondences between latent and reference images when making their determinations. This additional information could be incorporated into a psychometric model to better understand variation in examiner thresholds for making latent evaluation and source decisions.

In addition to research studies, forensic examiners also participate in annual proficiency tests, for which psychometric modeling can also be used (Luby and Kadane, 2018). While current proficiency tests are generally perceived to be easy with high-quality images (Gardner et al, 2020), they can be misinterpreted in legal contexts (Garrett and Mitchell, 2017). IRT-like models should be adopted for all proficiency testing. This would allow for the standardization of examiner scores across multiple years, adjusting for exams that were easier or harder than other exams. Research is also currently being conducted on blind proficiency tests, (see Mejia et al (2020) for overview), in which participants are unaware that they are being tested. This process is more complicated to implement than standard 'open' proficiency tests, as items need to be integrated within regular casework. Psychometrics could provide the methods for validating such tests and comparing results to open proficiency tests.

While we have focused on fingerprint identification throughout this paper, forensic science is a broad term used to describe many scientific fields, each of which relies at least partially on human decision-making. Psychometric models could be applied to each of these scientific areas to better understand the variability in examiner decision-making and potential impacts on final case outcomes.

Forensic science is an area ripe for psychometrics due to variation and uncertainty among forensic examiners, as well as the varying quality of evidence and corresponding difficulty in the analysis task. However, there are also challenges including privacy concerns for participants and for items, responses that are not keyed as correct or incorrect, and the sequential structure of forensic decision-making that must be accounted for. Through complex psychometric modeling, along with domain expertise, we can better understand variation in forensic decision-making and factors that impact that variation.

Bibliography

De Boeck P, Partchev I (2012) Irtrees: Tree-based item response models of the glmm family. Journal of Statistical Software, Code Snippets 48(1):1–28, DOI 10.18637/jss.v048.c01, URL https://www.jstatsoft.org/v048/c01

Dror IE, Langenburg G (2019) cannot decide: the fine line between appropriate inconclusive determinations versus unjustifiably deciding not to decide. Journal of forensic sciences 64(1):10–15

Friction Ridge Subcommittee of the Organization of Scientific Area Committees for Forensic Science (2017) Guideline for the Articulation of the Decision-Making Process Leading to an Expert Opinion of Source Identification in Friction Ridge Examinations. https://www.nist.gov/system/files/documents/2020/03/23/OSAC%20FRS%20ARTICULATION%20Document%20Template%202020_Final.pdf, online; accessed 15 September 2020

Friction Ridge Subcommittee of the Organization of Scientific Area Committees for Forensic Science (2019) Friction Ridge Process Map (Current Practice). https://www.nist.gov/system/files/documents/2019/12/10/Friction%20Ridge%20Process%20Map_December%202019.pdf, online; accessed 15 September 2020

Gardner BO, Kelley S, Pan KD (2020) Latent print proficiency testing: An examination of test respondents, test-taking procedures, and test characteristics. Journal of forensic sciences 65(2):450–457

Garrett BL, Mitchell G (2017) The proficiency of experts. U Pa L Rev 166:901

Jeon M, De Boeck P (2016) A generalized item response tree model for psychological assessments. Behavior research methods 48(3):1070–1085

Kerkhoff W, Stoel R, Berger C, Mattijssen E, Hermsen R, Smits N, Hardy H (2015) Design and results of an exploratory double blind testing program in firearms examination. Science & Justice 55(6):514 – 519, DOI https://doi.org/10.1016/j.scijus.2015.06.007

Luby A (2019) Accounting for individual differences among decision-makers with applications in forensic evidence evaluation. PhD thesis, Carnegie Mellon University. Available from: http://www.swarthmore.edu/NatSci/aluby1/files/luby-dissertation.pdf

Luby A, Mazumder A, Junker B (2020) Psychometric analysis of forensic examiner behavior. Behaviormetrika 47:355–384

Luby AS, Kadane JB (2018) Proficiency testing of fingerprint examiners with Bayesian Item Response Theory. Law, Probability and Risk 17(2):111–121

Mejia R, Cuellar M, Salyards J (2020) Implementing blind proficiency testing in forensic laboratories: Motivation, obstacles, and recommendations. Forensic science international Synergy 2:293298, DOI 10.1016/j.fsisyn.2020.09.002, URL https://europepmc.org/articles/PMC7552087

R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

Stan Development Team (2018a) RStan: the R interface to Stan. URL http://mc-stan.org/, r package version 2.18.2

Stan Development Team (2018b) Stan Modeling Language Users Guide and Reference Manual. URL http://mc-stan.org

Stern HS (2017) Statistical issues in forensic science. Annual Review of Statistics and Its Application 4:225–244

Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2011) Accuracy and reliability of forensic latent fingerprint decisions. Proceedings of the National Academy of Sciences 108(19):7733–7738

Ulery BT, Hicklin RA, Buscaglia J, Roberts MA (2012) Repeatability and reproducibility of decisions by latent fingerprint examiners. PloS one 7(3):e32,800

Ulery BT, Hicklin RA, Roberts MA, Buscaglia J (2014) Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. PloS one 9(11):e110,179