Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: www.sciencedirect.com/journal/forensic-science-international-synergy

A method for quantifying individual decision thresholds of latent print examiners

Amanda Luby

Department of Mathematics & Statistics, Swarthmore College, USA

ABSTRACT

In recent years, 'black box' studies in forensic science have emerged as the preferred way to provide information about the overall validity of forensic disciplines in practice. These studies provide aggregated error rates over many examiners and comparisons, but errors are not equally likely on all comparisons. Furthermore, inconclusive responses are common and vary across examiners and comparisons, but do not fit neatly into the error rate framework. This work introduces Item Response Theory (IRT) and variants for the forensic setting to account for these two issues. In the IRT framework, participant proficiency and item difficulty are estimated directly from the responses, which accounts for the different subsets of items that participants often answer. By incorporating a decision-tree framework into the model, inconclusive responses are treated as a distinct cognitive process, which allows inter-examiner differences to be estimated directly. The IRT-based model achieves superior predictive performance over standard logistic regression techniques, produces item effects that are consistent with common sense and prior work, and demonstrates that most of the variability among fingerprint examiner decisions occurs at the latent print evaluation stage and as a result of differing tendencies to make inconclusive decisions.

1. Introduction

In pattern evidence disciplines, final source decisions are often left to individual forensic examiners, who may have different internal thresholds for making decisions due to their training, experience, or visual acuity [1–3]. 'Black Box' error rate studies have emerged as one approach for quantifying the overall performance of these disciplines [4], but the resulting error rates, aggregated over many participants and items, ignore the systematic variability present. Errors are not equally likely across comparisons or examiners and have been concentrated among a subset of participants in forensic black box studies in fingerprints [5], palm prints [6], bullets and cartridge cases [7], and handwriting [8].

Furthermore, since the goal of these studies is to provide an estimate of casework error rates, it is important to include items that range from straightforward comparisons to very difficult comparisons that examiners may rarely encounter in casework. Since many items must be included to provide a fairly representative sample of casework tasks, participants are often given a random subset of items to analyze. For example, [5] included an item pool of 744 comparisons, which would be time and cost prohibitive for every participant to analyze. Instead, each participant was shown a subset of roughly 100 items. If different participants see different items, it is problematic to directly compare their individual error rates, since they may have seen items of varying difficulty.

Item Response Theory (IRT), a class of statistical models used extensively in educational testing, is an alternative approach for analyzing such data [9,10]. In the IRT framework, participant proficiency and item difficulty are assumed to be latent,[1] or unobserved, variables that govern the responses. Latent variables cannot be measured directly, but are inferred through observed variables using a mathematical model. For example, participant proficiency is a latent variable that is estimated using participant score on a given exam. Since both participant proficiency and item difficulty are estimated simultaneously from the data, participants who answer easier questions incorrectly are penalized more than participants who answer more difficult questions incorrectly.

Recently, the treatment of 'inconclusive' decisions has received much attention in forensic science, since such responses do not fit neatly into the error rate framework but are common in research studies and in casework [1,11–14]. Individual variability in inconclusive decisions has been demonstrated repeatedly in pattern evidence disciplines: latent prints [5,6], handwriting [8], and firearms [7,15] provide a few examples. IRT models are quite flexible and can be adapted for many different settings, including instances in which multiple internal decisions must be made. In this work, we use Item Response Trees [16,17] for the forensic science setting, which combine features of a decision tree with features from IRT, and provide robust individual propensities to make

[1] This usage differs from the usage of 'latent' in forensic science, or an impression that has been transferred to another surface. Both senses of the word are used in this paper, with the forensic concept including the word 'print' for clarity.

https://doi.org/10.1016/j.fsisyn.2023.100340
Received 31 March 2023; Received in revised form 30 May 2023; Accepted 13 June 2023
Available online 30 June 2023

inconclusive and no value decisions, in addition to traditional proficiency estimates.

IRT has recently shown promise in the fingerprint domain: [18] applied IRT models to annual proficiency test data and demonstrated the benefit of including harder items through simulation. [19] used an IRTree approach to develop an 'answer key' for inconclusive responses and measure variability in inconclusive tendencies, but did not distinguish between correct conclusive responses and errors. [20] used a latent variable model and found variability in the strength of evidence for different pairs of prints across two different studies, but did not model behavior at the individual examiner level. [2] posited that variability among examiners can be seen as the effect of implicit decision thresholds, which differ substantially across individuals, a phenomenon that has also been hypothesized in Ref. [1]. However, that approach used the observed frequencies of each decision to quantify these thresholds, which does not account for different participants responding to different subsets of questions. Furthermore, the frequency of inconclusive decisions may also depend on the quality of the latent print image, which has been observed in prior work but has yet to be formally incorporated as a predictor into a statistical model for forensic decision-making. In this paper, we present an IRTree model as one possible solution for combining information from multiple studies into a single model that distinguishes inconclusive from conclusive responses, and correct from incorrect responses. Using results from multiple studies in the latent fingerprint domain [5,21,22], we quantify the internal tendencies to make no value, inconclusive, and correct decisions, while accounting for the quality of the latent print (as calculated by the LQMetric software [23]). We confirm the findings from prior research using this model, present improvements and additional insights that can be gained from the IRT-based approach, and discuss how these types of models could be implemented in forensic science in the future.

In Section 2, we introduce the particular IRT and tree model that we apply to the [5] data in more detail. Section 3 presents the results from this analysis and provides comparisons to similar prior work Finally, Section 4 discusses implications of the results and future work in this area.

2. Methods

2.1. Item Response Theory

In order to fully specify the model, we first introduce the *Item Response Matrix* as a data representation scheme for results from error rate studies. For a study with I participants and J items, we represent the responses as an $I \times J$ matrix, $\boldsymbol{Y}$. If responses are scored as correct/incorrect, a 1 represents a correct response, a 0 represents an incorrect response, and a missing entry means that the participant was not assigned that item. Below is an example of such a matrix:

$$\mathbf{Y} = \begin{bmatrix} 1 & 0 & - & \dots & 1 \\ 0 & - & 1 & \dots & 0 \\ 1 & 1 & - & \dots & - \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & - & \dots & 1 \end{bmatrix}.$$

To obtain the participant scores, we can sum each row. To obtain the item scores, we can sum each column. In general, we expect high proficiency participants to have relatively high scores and high difficulty items to have relatively low item scores. In the IRT framework, instead of using the raw participant and item scores to assess proficiency and difficulty, we model the probability of a correct response for each entry in the matrix (e.g. $Y_{ij} = 1$). This approach assumes that participant proficiency is a latent variable, which is not directly observed but governs the observed responses. Similarly, item difficulty is also a latent variable that is estimated from the responses.

The Rasch Model [9,24] is a relatively simple, yet powerful, IRT model, and serves as the basis for the model presented in Section 2.2. The probability of a correct response is modeled as the logistic function of the difference between the participant proficiency, denoted $\theta_i (i = 1, \dots, I)$, and the item difficulty, denoted $b_j (j = 1, \dots, J)$:

$$P(Y_{ij} = 1) = \frac{1}{1 + \exp(-(\theta_i - b_j))}. \tag{1}$$

To identify the model, we use the convention of constraining the mean of the participant parameters (μ_θ) to be zero. This allows for an intuitive interpretation of both participant and item parameters relative to the "average participant". If $\theta_i > 0$, participant i is of "above average" proficiency and if $\theta_i < 0$, participant i is of "below average" proficiency. If $b_j = 0$, the model expects an average participant to make a correct decision 50% of the time. Similarly, if $b_j < 0$ item j is an "easier" item and the average participant is more likely to answer item j correctly. If $b_j > 0$ then item j is a more "difficult" item and the average participant is more likely to answer item j incorrectly. Other common conventions for identifying the model include setting a particular b_j or the mean of the b_js to be zero.

The item characteristic curve (ICC) describes the relationship between proficiency and performance on a particular item (see Fig. 1 for examples). For item parameters estimated under a Rasch model, all ICCs are standard logistic curves with different locations on the latent difficulty/proficiency scale.

In cases where participants answer different subsets of items, it is possible for high proficiency participants to have lower raw scores than lower proficiency participants. While somewhat non-intuitive, this phenomenon is actually a benefit of using the IRT-based approach: a participant who sees only very easy items and makes a couple of errors is penalized more than a participant who sees very difficult items and makes more errors.

The two-parameter logistic model (often referred to as the 2 PL) and three-parameter logistic model (3 PL) are additional popular IRT models [25]. They are both similar to the Rasch model in that the probability of a correct response depends on participant proficiency and item difficulty, but additional item parameters are also included. IRT models for outcomes with more than two categories are known as polytomous response models. Polytomous response data arises often in surveys where responses are collected with a Likert scale (i.e. strongly agree to strongly disagree) or when certain responses can be scored as partial credit. We omit a full discussion of these models here, but further reading may be found in Refs. [10,26]. We do not use the 2 PL and 3 PL models in this paper, since the number of responses per item is insufficient for estimating double (or triple) the number of item parameters. Although source decisions could be represented as a three-level ordinal scale (i.e. Exclusion < Inconclusive < Identification) we use an alternative approach which allows inconclusive responses to be governed by separate latent variables.

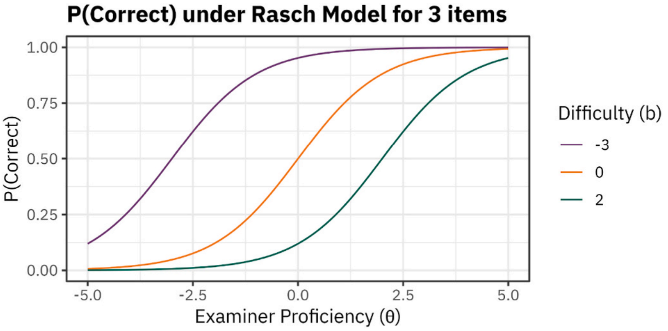


Fig. 1. Item Characteristic Curves (ICC) for 3 items of varying difficulty for the Rasch model. Examiners with higher proficiency (θ) have a higher probability of answering each question correctly than lower proficiency examiners.

2.2. Item Response Trees (IRTrees)

Item Response Trees (IRTrees, [16]) use decision trees to describe hypothesized cognitive processes, where each node corresponds to a possible sub-decision and the leaves represent observed outcomes. The IRTree formulation can represent a wide variety of response formats and response processes, easily adapted for binary responses, one-dimensional scales, bipolar scales, and Likert responses [17]. IRTrees have been used in general applications such as differentiating types of intelligence [27], response styles in multiple-choice items [28], and modeling answer change behavior [29]. In the forensic science setting, IRTrees have shown to be useful for representing sequential decision-making processes when an answer key does not exist [19].

Item Response Trees can be estimated using frequentist or Bayesian implementations. For the [5] data, we chose a Bayesian implementation due to the large variability in number of responses per item at different nodes in the tree. A Bayesian approach also provides a natural hierarchical structure for incorporating additional information about the participants and the items, as well as the uncertainty for all estimated parameters.

Fig. 2 shows the IRTree model formulation for the [5] study. For each participant × item pair, the first sub-decision (denoted Y_1^*) is whether the latent print was deemed to be *No Value* or not. Since participants were asked to make this decision before seeing the reference print, we believe it represents a distinct cognitive process from the remainder of the comparison task. The second sub-decision (Y_2^*) is whether the comparison was deemed to be *Inconclusive* or not. This outcome occurred when participants found the latent print to be suitable for comparison, but the comparison did not result in an *Identification*[2] or *Exclusion* decision. It is impossible to tell from the data alone whether inconclusive decisions arise from a different cognitive process than conclusive decisions, but separating inconclusive responses into a sub-decision allows for the quantification of individual participant and item tendencies to make an inconclusive decision. The study also recorded reasons for inconclusive decisions, which could provide further information, but we combine all responses into a single inconclusive response since the reasons given were often not repeated in follow-up study [21]. Finally, the third sub-decision (Y_3^*) represents whether the participant made a correct conclusive decision (making an identification on a same-source pair or an exclusion on a different-source pair) or an error (an exclusion on a same-source pair or an identification on a different-source pair). Since the no value and inconclusive decisions are modeled explicitly as distinct outcomes, we do not have to score them as correct or incorrect, avoiding the issues raised in, e.g., Refs. [11–14]. For the complete probabilities of observing a response in each category, see Appendix B.

$$\pi_{kij} = P(Y_{kij} * = 1) \tag{2}$$

$$\log\left(\frac{\pi_{kij}}{1-\pi_{kij}}\right) = \theta_{ki} - b_{kj} \tag{3}$$

The participant and item tendencies estimated at node Y_3^* (θ_{3i} and b_{3j}) are the same proficiency and difficulty estimates as introduced in 2.1. The participant tendency estimated at node Y_1^*, θ_{1i}, is participant i's tendency to make 'no value' decisions: participants with $\theta_{1i} > 0$ are more likely to make 'no value' decisions than average and participants with $\theta_{1i} < 0$ are less likely to make 'no value' decisions than average, after accounting for the items that each participant saw. At node Y_2^*, θ_{2i}, is participant i's tendency to rate items as inconclusive, conditional on a 'has value' decision at node Y_1^*. Again, positive estimates indicate that participant i is more likely to be inconclusive, while negative estimates indicate that participant i is less likely to be inconclusive. On the item side, negative b_{1j}'s indicate items that are more likely to be rated as 'no value' by the average examiner, and positive b_{1j}'s indicate items that are more likely to be rated as 'has value' by the average examiner. Similarly, negative b_{2j}'s indicate items that are more likely to be rated as inconclusive by the average examiner, and positive b_{2j}'s indicate items that are more likely to be rated as conclusive by the average examiner.

The tree structure in 2 is one of many possible decision tree structures. [19] provides a few other tree structures that (a) incorporate the reason for inconclusive decisions, and (b) do not distinguish between correct conclusive decisions and erroneous conclusive decisions. The tree structure presented here allows us to assess how *No Value* and *Inconclusive* tendencies are related to participant proficiency and item difficulty. It's also important to note that the current model does not explicitly distinguish between false positive and false negative errors. Since only six false positive errors were observed in the study, separating those decisions into a separate branch of the tree would not lead to meaningful estimates. Instead, we incorporate ground truth of the item into an explanatory model for the item tendencies.

In addition to the structure of the tree defined above, we can also incorporate additional explanatory information about the latent/reference print pairs into the item estimates. For example, since false negative errors are far more likely than false positive errors, we might expect the item tendencies at node Y_3^* (b_{3i}) to vary depending on whether the item is a same-source or different-source pair. We might also expect that item tendencies for no value (Y_1^*) and inconclusive (Y_2^*) decisions to be related to the image quality of the latent print. Equation (4) shows the linear model used to incorporate these two variables, where k indexes the node in the IRTree ($k = 1, 2, 3$) and j indexes the item ($j = 1, \ldots, J$). An interaction term between Mated and LQM is included, which allows the relationship between item tendency and LQM to vary depending on whether the pair is same-source or not. Since an ε term is included for each item estimate, we also allow for the possibility that other factors impact the item tendencies that are not captured by the two included variables. Items with especially large ε values might indicate latent/reference print pairs that should be further investigated.

$$b_{kj} = \beta_{k0} + \beta_{k1} \times \mathrm{Mated}_j + \beta_{k2} \times \mathrm{LQM}_j + \beta_{k3} \times \mathrm{Mated}_j \times \mathrm{LQM}_j + \varepsilon_{kj} \tag{4}$$

Positive $\beta_{k,m}$ estimates increase the item tendency towards the right branch at each node of the IRTree. For example, if the same source items are indeed more likely to be incorrect at node Y_3^*, we would expect a positive β_{31} estimate. If high-LQM items are more likely to be rated as 'has value', we would expect a positive β_{12} estimate. We might also expect examiners to require higher quality latent prints to make conclusive decisions on same-source prints compared to different-source prints. This would manifest as a positive β_{23} estimate.

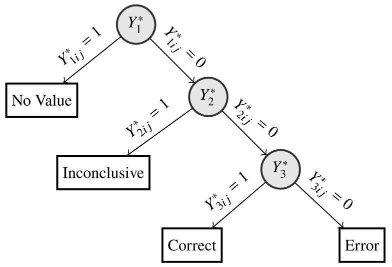
Because we implement a Bayesian approach to estimation, prior distributions must be assigned to all parameters in order to obtain posterior estimates. In general, we choose weakly informative prior distributions, which result in a more efficient sampling procedure without the need to make strong assumptions. For all prior distribution formulations and details on model fitting, see Appendix B.

3. Results

3.1. Overall prediction accuracy

The first metric we use to evaluate the performance of the model is simple prediction accuracy: how often does the model predict the correct outcome for each participant × item pair? Table 1 shows the misclassification rates for the IRTree model on *no value* decisions, *identification* decisions on mated pairs, *exclusion* decisions for non-mated pairs, and *inconclusive* decisions. As a reference point, the logistic regression models in Ref. [2] that incorporated both participant and

[2] In the [5] study, a same-source conclusion was recorded as an *Individualization*, but we use *Identification* here to be consistent with current reporting recommendations [30].



$$\pi_{kij} = P(Y_{kij}* = 1) \tag{2}$$

$$\log(\frac{\pi_{kij}}{1 - \pi_{kij}}) = \theta_{ki} - b_{kj} \tag{3}$$

Fig. 2. Example IRTree model tree structure for the [5] data (left) and probability formulations for each node k. For alternative model structures, see Ref. [19].

Table 1
Misclassification rates for the original error rate study using the logistic models from Ref. [2]; which use empirical outcome rates for each examiner (E%) and item (L%) as predictors, and the IRTree model presented in Fig. 2. The IRTree model achieves an improvement over the logistic models for each outcome, demonstrating the predictive benefits of a latent variable approach.

Outcome	Logistic Models	IRTree Model
	P(Outcome |E%, L%)	P(Outcome |$\boldsymbol{\theta}$, $\boldsymbol{b}$)
No Value	14%	10.6%
Identification (Mated)	18%	8.3%
Exclusion (Non-mated)	15%	4%
Inconclusive	–	20.1%

item rates are shown in the first column. The primary differences between the logistic models and the IRTree model are (1) the logistic models use observed decision rates as predictors, while the IRTree model uses latent variables, and (2) the logistic models for IDs and Exclusions are fit using subsets of the data, while the IRTree is fit to the full data but conditions on prior decisions. For example, the logistic regression model for *No Value* decisions from Ref. [2] uses the same outcome and data as node Y_1 from the IRTree model, the only difference is that the predictors are empirical rates instead of latent variables. The other two logistic regression models used for comparison (labeled 'ID' and 'Exclusion' in 1) each take a subset of the data (ground truth matches and non-matches, respectively), while the IRTree model predicts correct responses conditional on a conclusive decision.

In all categories for which a comparison is available, the IRTree model achieves an improvement over the separate logistic models, ranging from 3.5 to 10%. For example, the logistic models using empirical outcome rates as predictors have a 15% misclassification rate on exclusion decisions among all non-mated pairs, while the IRTree model achieves only a 4% misclassification rate. Since the logistic models used the observed rates for each decision, this improvement demonstrates the benefits of (1) a latent variable approach, which accounts for the different item sets shown to each participant, and (2) employing a single model to more efficiently share information, rather than different models for each possible outcome.

3.2. *Posterior predictive check*

When fitting Bayesian models, posterior predictive checks should be done in order to ensure that the resulting estimates are consistent with the observed data [31,32]. In our case, a primary quantity of interest is the response distribution for each examiner. Fig. 3 shows 95% posterior prediction intervals for the number of responses in each category for each participant across the entire study compared to the actual number of responses that were observed, over 1000 MCMC samples. The diagonal line represents a perfect prediction. 169 (100%) of the intervals for No Value and Inconclusive responses contain the observed frequency, 168 (99.4%) of the intervals for Correct contain the observed frequency, and 127 (75.1%) of the intervals for Errors contain the observed frequency. Since most posterior intervals overlap with the diagonal line in all four panels, the model is generally able to predict the overall frequencies for each participant and is consistent with the observed data, lending credibility to the fit of the model. However, we should note that the model performance for Errors is noticeably weaker than for the other outcomes, suggesting that errors may not be fully explained by examiner proficiency and item difficulty.

3.3. *Out-of-sample retest performance*

The [21] study, which asked participants to re-examine a subset of comparisons that they had already completed, provides a unique

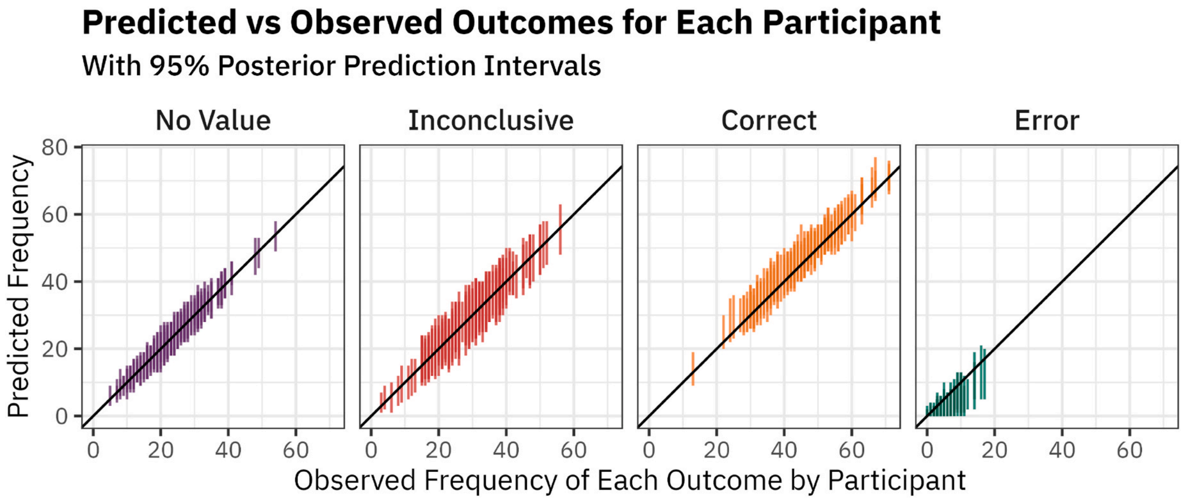


Fig. 3. Posterior predictions for the number of 'no value', 'inconclusive', 'correct' and 'error' responses for each participant based on the IRTree model in 2. The diagonal line represents a perfect prediction. Since most intervals intersect the diagonal line, model performance is strong at the participant level.

opportunity for model validation. Using the model estimated from the original study, we can see whether it is able to predict new observations for the same examiner × item pair on the retest. This is an example of evaluating the ‘out-of-sample’ predictive performance: the retest data was not used to estimate the model and therefore provides a more realistic setting for evaluating how well the model will predict new observations. If the model is not able to predict responses on the retest, it may suggest overfitting to the original data.

Table 2 shows the overall accuracy of the IRTree model and the logistic models on the out-of-sample retest data, illustrating the robustness of the IRTree model compared to the logistic models. Both models clearly perform worse on the retest data than they do on the original data, but that is to be somewhat expected since only 73.5% of decisions (No Value, Inconclusive, Correct, or Error) were repeated by participants. The IRTree model substantially outperforms the logistic models based on observed decision rates, particularly on how often it correctly predicts ID decisions on mated pairs. Because the logistic models are based on the empirical decision rates of each examiner, they are more sensitive to differing item sets than the IRTree model. In the original study, participants were assigned item subsets consisting of 50–75% same-source pairs. On the retest, some examiners were assigned *only* same-source pairs, and the logistic model will substantially under-predict how many IDs are made on the retest. Since the predictors in the logistic models are based on observed decision rates, it is harder to separate examiner tendencies from item tendencies. The latent variable based approach of the IRTree model is more robust to differing base rates and small changes in internal decision-making across time, making it more useful for predicting decisions on new data.

Fig. 4 shows the out-of-sample posterior prediction intervals for the retest data for each examiner on each category of responses, with the diagonal reference line representing perfect prediction. In general, the posterior predictions are worse than for the original data, but since most intervals overlap with the reference line, the model is performing reasonably well. 82 (75.9%) of the No Value intervals contain the observed frequency; 91 (84.2%) of the Inconclusive intervals contain the observed frequency; 93 (86.1%) of the intervals for Correct contain the observed frequency, and 94 (87%) of the intervals for Error contain the observed frequency. Of particular interest are the intervals that *don’t* overlap with the reference line: these represent participants that may have changed their decision-making thresholds between the initial study and the retest.

3.4. Correlation between latent parameters

For participants, we estimate a positive correlation between θ_1 and θ_2 ($\hat{\sigma}_{1,2} = 0.36$, [0.22, 0.495]). Therefore, participants who are more likely to make ‘no value’ decisions are also mildly more likely to make ‘inconclusive’ decisions, and vice versa. A negative correlation between θ_3 and both θ_2 ($\hat{\sigma}_{2,3} = -0.24$, [$-0.40$, -0.05]) and θ_1 ($\hat{\sigma}_{1,3} = -0.25$, [$-0.43$, -0.04]) was estimated, meaning that participants who are more likely to make correct conclusive decisions are slightly less likely to make inconclusive and no value decisions. Distributions of the θ point estimates, sample correlations, and a scatterplot matrix can be found in Appendix C.

On the item side, we estimate the relationship between parameters to have the same direction but stronger magnitude. Items that are likely to be rated as ‘no value’ are more likely to be rated as inconclusive ($\hat{\sigma}_{1,2} = 0.71$, [0.64, 0.77]) but tend to be less likely to be correct if a conclusive decision is reached ($\hat{\sigma}_{1,3} = -.74$, [$-0.81$, -0.65]). Items that tend to be rated as inconclusive also tend to be less likely to be correct if a conclusive decision is reached ($\hat{\sigma}_{2,3} = -.81$, [$-0.86$, -0.74]).

Table 2
Misclassification rates for decisions on the retest data using the logistic models from Ref. [2] and the IRTree model.

	Logistic Models	IRTree Model
	P(Outcome |E%, L%)	P(Outcome |$\boldsymbol{\theta}$, $\boldsymbol{b}$)
VID	35.7%	18.5%
ID (Mated)	75.4%	14%
Excl (Non-mated)	28.2%	6.4%
Inconclusive	–	27.8%

The sample correlations, distributions of the b point estimates, and a scatterplot matrix are shown in Fig. 5. There are a few notable trends. First, when separating items by ground truth, the sample correlations of the point estimates become even more pronounced. Second, since the false positive rate was so low, we would expect the same-source items to be more difficult than the different source items, and indeed we see that all positive b_3 items are same-source items. Finally, we can see the relationship between LQ Metric and ground truth on the item estimates: there are more same-source items than different-source items, the same-source items have proportionally more low quality latent prints, and there are more same-source items with negative b_1 and b_2 estimates. These relationships are discussed more formally in Section 3.6.

While these relationships should not be surprising, they do lend credibility to the parameter estimates. These results also provide some quantification of the relative predictive weight of participant versus item tendencies. The correlations between participant parameters are weaker in magnitude, suggesting that while there is some relationship between tendencies at different points in the tree, there is also a high amount of variability. On the other hand, the relationships on the item side are estimated to be relatively strong.

3.5. Examiner differences

Since we estimate an examiner tendency, θ, for each examiner at each node in the decision tree, we are able to quantify the impact of internal decision thresholds. Fig. 6 shows how the probability of making a decision (No Value, Inconclusive, or Correct) changes as θ_k increases. Each panel corresponds to a decision, and each line corresponds to a hypothetical item based on the percentiles of b estimates (10th, 25th, 50th, 75th, and 90th) for that decision. For an item where $\theta_1 = 0$, we can see that the probability of making a no value decision ranges from near zero for the 90th, 75th, and 50th percentile of items, near 0.3 for the 25th percentile of items, and near 1 for the 10th percentile of items.

Note that each panel covers different x-axis ranges to ensure that the full range of possible θ values are shown. The ranges for θ_1 and θ_2 are larger than for θ_3. Similarly, the b values also cover different ranges according to the decision that they correspond to, although we display only the percentile for simplicity.

The difference between examiners is most noticeable when making No Value and Inconclusive decisions, since that is where we observe the greatest differences in probability for a given item. This means that an examiner at the lower end of the θ spectrum is likely to make a different decision than an examiner at the upper end of the θ spectrum for the same item. On the other hand, there is little variation in the third decision. For any given item, examiners are likely to agree regardless of where they lie on the θ spectrum. In practical terms, this means that examiners are largely likely to agree on identification/exclusion decisions, but much more likely to disagree for no value or inconclusive decisions.

For all decisions, the differences between examiners are most pronounced when b is near zero. When b is near zero, an average examiner is equally likely to make either decision (e.g. ‘no value’ versus ‘has value’ or ‘inconclusive’ versus ‘conclusive’). As b gets further from zero, the differences between examiners becomes more negligible, meaning most examiners will agree on latent prints that are very clearly “no value” or “inconclusive”.

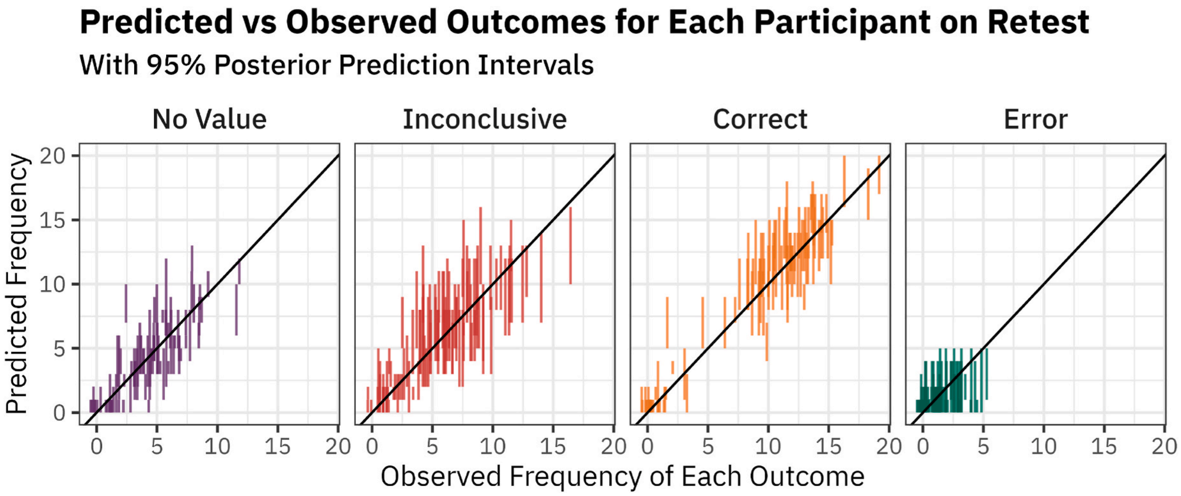


Fig. 4. Predictions for the number of 'no value', 'inconclusive', 'correct' and 'error' responses for each participant on a retest [21], using the posterior estimates from the original model. While performance is not as strong as the original study, most intervals overlap with the diagonal lines, suggesting reasonable out-of-sample performance by the IRTree model.

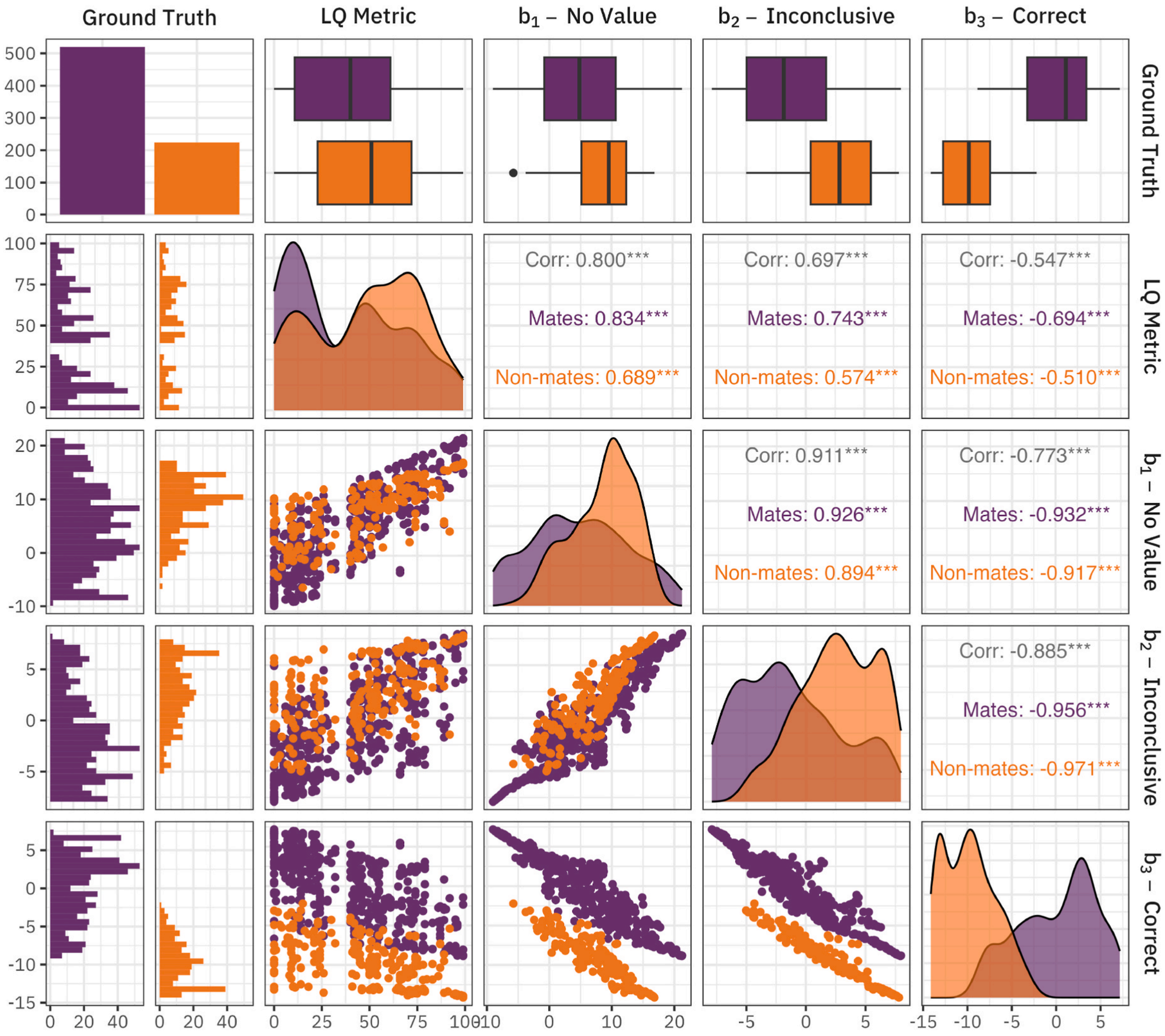


Fig. 5. Scatterplots and sample correlations of b_1, b_2, and b_3 point estimates; along with ground truth and LQ Metric for each item. Note that the correlations here are the sample correlations of the point estimates, and are larger in magnitude than the population estimates reported.

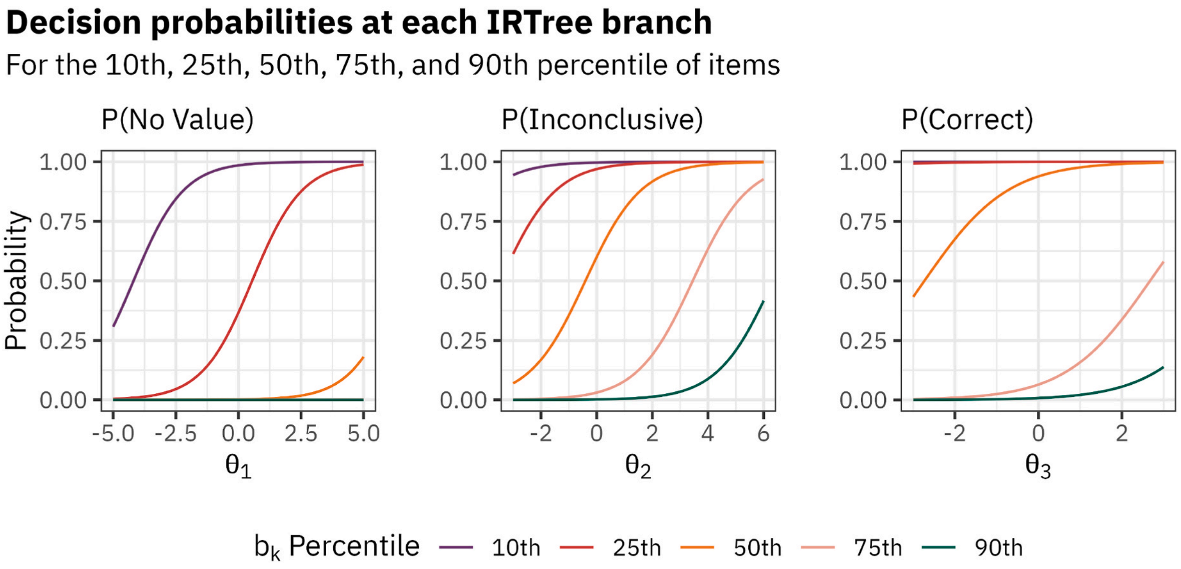


Fig. 6. Decision probabilities at each node of the IRTree for all values of θ and the 10th, 25th, 50th, 75th, and 90th percentile of items.

3.6. Effects of image quality and ground truth

Since image quality (LQM values) and ground truth (mated or non-mated) were incorporated into the models as predictors for the item tendencies (as in Equation (4)), we can draw conclusions about how LQM and ground truth impact the probability of a No Value, Inconclusive, and Correct decision. The posterior estimates and 95% posterior intervals for each β coefficient at each branch in the IRTree are shown in Table 3. Most estimates are measurably different than zero, suggesting that both ground truth and LQM impact item tendencies at each node in the IRTree. However, an interaction term was included (β_{k3}) which makes it difficult to draw conclusions based on the coefficients alone.

To provide a more intuitive interpretation of the coefficients, we can examine the marginal effects plot. For all possible combinations of LQM and Ground Truth, we compute the predicted b_k based on the linear model in Equation (4) and estimates from Table 3 (these predictions on the item tendency scale are shown in Appendix D). We then compute P (No Value), P(Inconclusive), and P(Correct) based on Equation (2), assuming a hypothetical average examiner ($\theta_k = 0$). The uncertainty intervals for the predicted probabilities are based on the coefficient intervals from Table 3. Fig. 7 shows the marginal effect of each predictor (*Ground Truth* and *Latent Quality*) on the probability of making a No Value, Inconclusive, or Correct decision. The right-most panel (*P(Correct)*) corresponds to the probability of making a correct conclusive decision at Node Y_3^*, given that neither a no value or inconclusive was reached. False positive errors are very rare, and so the probability of a correct decision for Non-Mates is high for all values of Latent Quality.

Table 3
Estimates and 95% posterior intervals for the coefficients in Equation (4).

Node	Coefficient	Estimate	Interval
Y_1^*	β_0	2.62	[1.23,4.04]
Y_1^*	β_1	−5.89	[-7.49,-4.33]
Y_1^*	β_2	12.29	[9.27,15.46]
Y_1^*	β_3	9.47	[5.83,13.18]
Y_2^*	β_0	−0.47	[-1.36,0.41]
Y_2^*	β_1	−4.73	[-5.80,-3.71]
Y_2^*	β_2	6.55	[4.95,8.2]
Y_2^*	β_3	4.14	[2.22,6.11]
Y_3^*	β_0	−7.13	[-9.35,-5.11]
Y_3^*	β_1	11.09	[8.84,13.54]
Y_3^*	β_2	−5.35	[-9.59,-1.33]
Y_3^*	β_3	−4.87	[-9.19,-0.53]

The probability of a correct decision for Mated pairs is low for low values of Latent Quality, but rapidly increases as the quality of the print increases, reaching a predicted 75% chance of being correct for a latent print with a Latent Quality of 50. The uncertainty intervals are quite broad; suggesting there is additional variability in item difficulty that is not explained by ground truth and latent print quality.

The middle panel (*P(Inconclusive)*) corresponds to the probability of making an inconclusive decision (if the latent print was deemed to be of value). As Latent Quality increases, the probability of observing an inconclusive decision decreases for both Mates and Non-Mates. We might expect fewer inconclusives on low-quality non-mated pairs compared to mated pairs, since the amount of information needed to make an exclusion is generally less than an identification (e.g., it is possible to exclude based on overall pattern instead of minutiae). Indeed it does appear that non-mated pairs are generally less likely to be inconclusive than mated pairs. However, the uncertainty intervals are broad enough that we cannot confirm a significant difference except for extremely low LQM values. We do observe a negative $\beta_{2,1}$ (−4.73 [−5.80,-3.71]), suggesting that mated pairs are generally more likely to be inconclusive than non-mated pairs. We also observe a positive $\beta_{2,3}$ (4.14 [2.22, 6.11]), suggesting that the relationship between LQ Metric and inconclusive tendency is different for mated versus non-mated pairs.

The left-most panel shows the marginal effect of Latent Quality and Ground Truth on the probability of coming to a 'No Value' decision. Since this decision occurs *before* the participant sees the reference print, Ground Truth should not impact the decision. However, the coefficients at this node for both Ground Truth and LQ Metric are significantly different than zero, and the marginal effect plot displays a very low P(No Value) for all LQM values in non-mated pairs. This could be due to proportionally more mated pairs with very low LQM values (see Fig. 5). Additionally, as latent print quality increases, P(No Value) tends to decrease at a faster rate than P(Inconclusive). For example, a latent print in a mated pair with latent quality of 25 results in a P(No Value) of about 12.5%, while the same latent print on average has over a 90% chance of being rated as inconclusive. This suggests that, on average, examiners may proceed with a comparison on low-quality prints for which an inconclusive decision is ultimately likely.

It is important to note that the coefficients presented here are based on the average impact on item tendencies across all item pairs and examiner decisions, and should not be interpreted as a deterministic relationship. That is, taking a latent print from a non-mated pair with a certain LQM and instead pairing it with a reference print taken from the same source as the latent will not necessarily increase the probability of

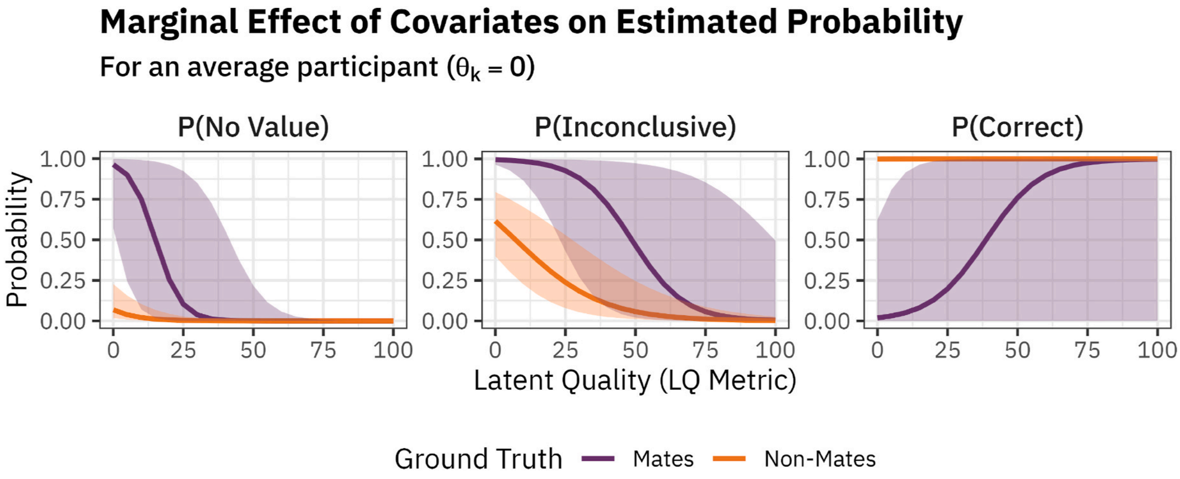


Fig. 7. Marginal effects of item covariates (Latent Quality and Ground Truth) on sub-decision probabilities for a hypothetical average examiner (with 95% posterior prediction intervals).

an inconclusive or decrease the probability of a correct response. These trends instead capture the average relationship between LQM and ground truth on item tendencies in this particular study. There was substantial variability in item tendencies even after accounting for these two predictors (see Appendix D). This variability could be due to additional characteristics of the latent prints (e.g. whether or not a delta was observed or certain minutiae configurations), image quality aspects that are not captured in the LQM, or simply inherent randomness in fingerprint comparisons. Since only LQ Metric and ground truth are available, we cannot know if there are other features that may explain the additional variability.

4. Discussion

In this paper, an IRT-based analysis was performed on the FBI “Black Box” latent print study. This statistical approach combined information from multiple studies on error rates [5], repeatability [21], and latent print quality [22] into a single model. Many of the same conclusions are drawn as prior work (e.g. Refs. [2,19], but the framework presented here estimates participant tendencies that account for the different subsets of items that each participant was shown. The latent variable approach presented here results in quantities that are directly comparable to one another and more robust to variation in item sets than the usual percent-observed approach.

[4] put forth forensic ‘black box’ studies as the gold standard to establish foundational validity for feature comparison methods, and black box studies have since been performed in a variety of forensic disciplines. However, such studies are difficult to design and expensive to run, since they require providing a large number of examiners with a substantial number of test items that are representative of those in casework. Furthermore, the results of these studies often emphasize only a few quantities (e.g., the overall false positive or false negative rate), which has been criticized for ignoring the variability present across examiners and items [33,34].

By incorporating analytical methods from psychometrics and educational testing, it is possible to extract much more information out of a black box study in addition to aggregated error rates. Using Item Response Theory, it is possible to obtain performance metrics specific to each comparison and participant in the study. Importantly, since these metrics are estimated with a latent variable approach, they account for different participants analyzing different subsets of comparisons.

A tree-based analysis provides a fuller picture of the range of decisions that may be expected for a given examiner or item and is particularly useful for this type of data where clear “correct” and “incorrect” responses do not exist (for example, on low quality images for which a no value or inconclusive is likely). ‘Inconclusive’ or ‘No Value’ decisions are not equally likely on every latent print comparison, and the methods presented here provide a rigorous way to quantify those tendencies. We have also demonstrated how this method can be used to explicitly quantify different internal decision-making thresholds among examiners through their latent variable estimates for ‘Inconclusive’ and ‘No Value’ decisions. In traditional proficiency tests, where the ground truth is known and image quality is high, tree-based analyses may not be necessary if ‘no value’ and ‘inconclusive’ decisions are not expected.

Tree-based approaches also provide a better sense of the variability across participants and comparisons. Combining decisions into a single model allows for more efficient borrowing of information and the explicit estimation of covariance among item and participant tendencies, demonstrating a clear relationship between ‘no value’, ‘inconclusive’, and ‘correct’ decisions, particularly at the item level.

In this paper, one decision tree structure is presented using a logistic probability model, fit within a Bayesian framework using one set of prior distributions. The model fit and posterior prediction checks in Section 3 lend credibility to these choices; but alternative choices are also possible. For example, logistic functions represent a small subset of possible probability mapping functions, and there are a number of parametric and nonparametric alternatives. Other functions, tree structures, or prior distributions could perform as well or better and lead to different interpretations.

These models require a large amount of data relative to most forensic datasets. The [5] study included 169 participants and an item bank of 744 latent/reference fingerprint pairs. Each participant was assigned roughly 100 items, resulting in 20–25 observations per item. The IRT-based analysis presented here produced estimates that were distinguishable from one another, but the associated uncertainty may be too large for some settings. Additional observations per item would lead to more precise estimates but would require either a smaller number of items (which may introduce additional sampling error) or a larger number of participants (which may be time or cost prohibitive).

Future work also includes exploring the possibility of implementing this method on training materials. In such a setting, personalized training and feedback could be provided, and no value and inconclusive decisions could be calibrated to be consistent with practicing latent print examiners. With a sufficiently large item bank, trainees could participate at multiple time points to demonstrate improved performance and more calibrated decisions.

IRT-based approaches show significant promise for annual proficiency tests and future large-scale error rate studies across forensic disciplines. While we have applied this method to latent fingerprint analysis, it can be easily adapted for other feature comparison domains such as palmar prints, firearms, handwriting, or shoeprints. Measuring and understanding variability in forensic decisions is the first step towards minimizing said variability, and IRT provides a set of well-developed statistical methods, theory, and tools to do so.

Funding acknowledgments

This work was funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

APPENDIX A. Software Acknowledgments

All analyses were done using R [35] and RStudio [36]. Data work was performed with the tidyverse packages [37], plots were made with ggplot [38], ggally [39], and patchwork packages [40], and all models were fit with the rstan package [41].

APPENDIX B. Bayesian Model Formulation

The probabilities for each outcome are computed as a product of Rasch models, conditional on the previous nodes in the tree:

$$\begin{aligned}
P(Y_{ij} = \text{No Value}) &= \text{logit}^{-1}(\theta_{i1} - b_{j1}) \\
P(Y_{ij} = \text{Inconclusive}) &= [1 - \text{logit}^{-1}(\theta_{i1} - b_{j1})] \times \text{logit}^{-1}(\theta_{i2} - b_{j2}) \\
P(Y_{ij} = \text{Correct}) &= [1 - \text{logit}^{-1}(\theta_{i1} - b_{j1})] \times [1 - \text{logit}^{-1}(\theta_{i2} - b_{j2})] \times \text{logit}^{-1}(\theta_{i3} - b_{j3}) \\
P(Y_{ij} = \text{Error}) &= [1 - \text{logit}^{-1}(\theta_{i1} - b_{j1})] \times [1 - \text{logit}^{-1}(\theta_{i2} - b_{j2})] \times [1 - \text{logit}^{-1}(\theta_{i3} - b_{j3})].
\end{aligned}$$

We fit this model under the Bayesian framework with Stan in R [35,41], using the following prior distributions,

$$\left.\begin{array}{cc}
\boldsymbol{\theta}_i & \overset{iid}{\sim} MVN_5(\mathbf{0}, \boldsymbol{\sigma_\theta} L_\theta L_\theta^{'} \boldsymbol{\sigma_\theta}) \\
\boldsymbol{b}_j & \overset{iid}{\sim} MVN_5(\boldsymbol{\beta}\mathscr{X}_j, \boldsymbol{\sigma_b} L_b L_b^{'} \boldsymbol{\sigma_b}) \\
L_\theta & \sim LKJ(4) \\
L_b & \sim LKJ(4) \\
\sigma_{k\theta} & \overset{iid}{\sim} \text{Half} - \text{Cauchy}(0, 2.5) k = 1, 2, 3 \\
\sigma_{kb} & \overset{iid}{\sim} \text{Half} - \text{Cauchy}(0, 2.5) k = 1, \ldots, 5 \\
\beta_{0k}, \beta_{1k}, \beta_{2k} & \overset{iid}{\sim} N(0, 5) k = 1, 2, 3
\end{array}\right\} \tag{5}$$

Here $\mathscr{X}_j$ is the column vector $(1, X_j)^{'}$, $\boldsymbol{\beta} = (\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_4})$ is the 3×4 matrix whose k^{th} row is $(\beta_{0k}, \beta_{1k}, \beta_{2k}, \beta_{3k})$, and $\boldsymbol{\sigma_b}$ is a 3×3 diagonal matrix with σ_{1b}, σ_{2b}, σ_{3b} as the diagonal entries; $\boldsymbol{\sigma_\theta}$ in the previous line is defined similarly. Multivariate normal distributions for $\boldsymbol{\theta}_i$ and $\boldsymbol{b}_j$ were chosen to estimate covariance between latent variables explicitly. 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. The recommended priors [42] for L and σ were used: an LKJ prior [43]; LKJ = last initials of authors) with shape parameter 4, which results in correlation matrices that mildly concentrate around the identity matrix ($LKJ(1)$ results in uniformly sampled correlation matrices), and half-Cauchy priors on σ_{kb} and $\sigma_{k\theta}$ to weakly inform the correlations. $N(0, 5)$ priors were assigned to the linear regression coefficients (β_k).

The complete Stan code for fitting the model is below.

```
data {
  int<lower=1> I;               // # items
  int<lower=1> J;               // # participants
  int<lower=1> N1;              // # node 1 sample size
  int<lower=1> N2;               // # node 2 sample size
  int<lower=1> N3;               // # node 3 sample size
  int<lower=1, upper=I> ii1[N1];  // vector of item IDs for node 1
  int<lower=1, upper=I> ii2[N2];  // vector of item IDs for node 2
  int<lower=1, upper=I> ii3[N3];  // vector of item IDs for node 3
  int<lower=1, upper=J> jj1[N1];  // vector of person IDs for node 1
  int<lower=1, upper=J> jj2[N2];  // vector of person IDs for node 2
  int<lower=1, upper=J> jj3[N3];  // vector of person IDs for node 3
  int<lower=0, upper=1> y1[N1];   // Node 1: 1 if No Value, 0 otherwise
  int<lower=0, upper=1> y2[N2];   // Node 2: 1 if Inconclusive, 0 otherwise
  int<lower=0, upper=1> y3[N3];   // Node 3: 1 if Correct, 0 otherwise
  int<lower=1> K;               // # splits in tree
  matrix[I,4] V;                // item covariates (with intercept,interaction)
}
parameters {
  matrix[K,I] b_tilde;
  matrix[K,J] theta_tilde;
  cholesky_factor_corr[K] L_OmegaT;
  vector<lower=0>[K] sigmaT;
  cholesky_factor_corr[K] L_OmegaB;
  vector<lower=0>[K] sigmaB;
  matrix[4, K] beta;
}
```

```
transformed parameters {
  matrix[I, K] b;
  matrix[J, K] theta;
  b = (diag_pre_multiply(sigmaB, L_OmegaB) * b_tilde)';
  theta = (diag_pre_multiply(sigmaT, L_OmegaT) * theta_tilde)';
}
model {
  L_OmegaB ~ lkj_corr_cholesky(4);
  sigmaB ~ cauchy(0, 2.5);
  to_vector(b_tilde) ~ normal(to_vector((V*beta)'), 1);
  L_OmegaT ~ lkj_corr_cholesky(4);
  sigmaT ~ cauchy(0, 2.5);
  to_vector(theta_tilde) ~ normal(0,1);
  to_vector(beta) ~ normal(0,5);
  y1 ~ bernoulli_logit(theta[jj1,1] - b[ii1,1]);
  y2 ~ bernoulli_logit(theta[jj2,2] - b[ii2,2]);
  y3 ~ bernoulli_logit(theta[jj3,3] - b[ii3,3]);
}
generated quantities {
  corr_matrix[K] OmegaB;
  corr_matrix[K] OmegaT;
  OmegaB = L_OmegaB * L_OmegaB';
  OmegaT = L_OmegaT * L_OmegaT';
}
```

. (*continued*).

APPENDIX C. θ Estimates

Figure 8 shows the distribution of the θ_k point estimates on the diagonal panels, pairs scatterplots of the θ_k point estimates on the lower panels, and sample correlations on the upper panels. Note that the correlations in the figure are the sample correlations of the point estimates, and are larger in magnitude than the population estimates reported in Section 3.4.

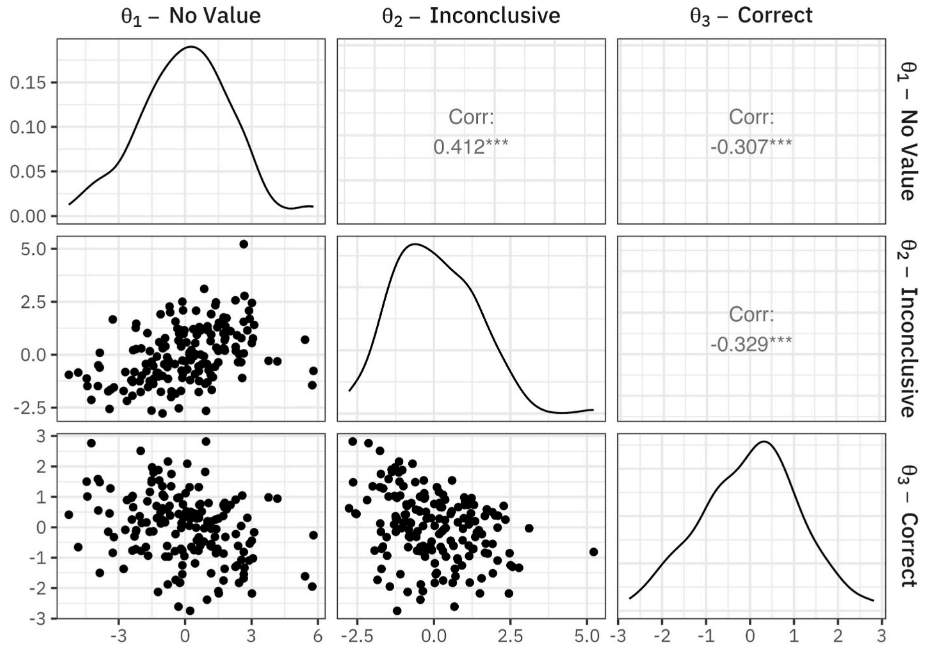


Fig. 8. Scatterplots and sample correlations of θ_1, θ_2, and θ_3 point estimates.

APPENDIX D. Predicted and Actual *b* based on LQ Metric and Ground Truth

Figure 9 shows the predicted b_k estimates based on Equation (4) and Table 3, along with the actual b_k point estimates.

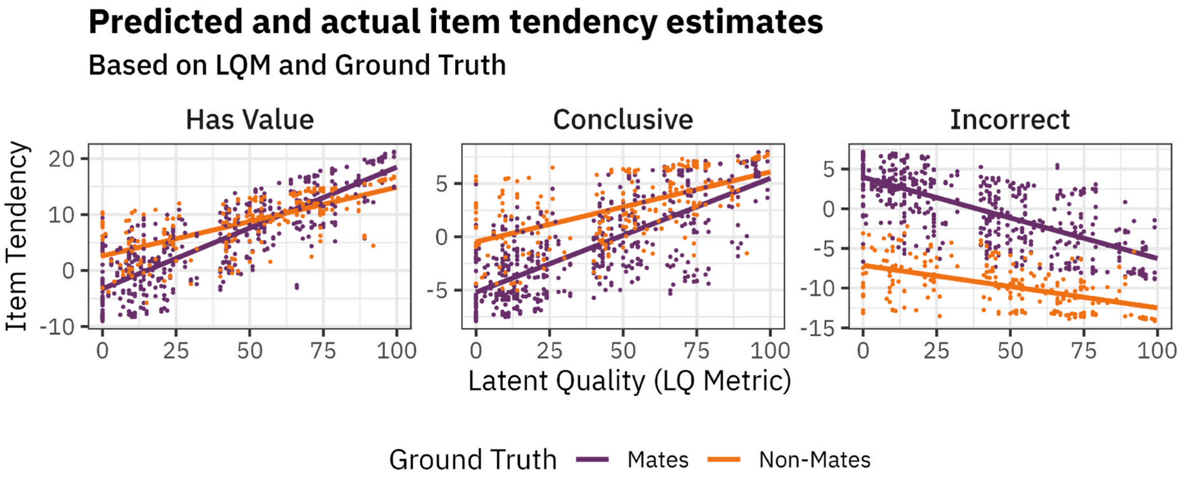


Fig. 9. Predicted and actual *b* estimates based on LQ Metric and Ground Truth (as in Equation (4))

References

[1] I.E. Dror, G. Langenburg, “cannot decide”: the fine line between appropriate inconclusive determinations versus unjustifiably deciding not to decide, J. Forensic Sci. 64 (1) (2019) 10–15.

[2] R.A. Hicklin, B.T. Ulery, M. Ausdemore, J. Buscaglia, Why do latent fingerprint examiners differ in their conclusions? Forensic Sci. Int. 316 (2020), 110542.

[3] B. Growns, J.D. Dunn, E.J. Mattijssen, A. Quigley-McBride, A. Towler, Match me if you can: evidence for a domain-general visual comparison ability, Psychon. Bull. Rev. (2022) 1–16.

[4] P. Pcast, Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Executive Office of the President of the United States, President’s Council, 2016.

[5] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. USA 108 (19) (2011) 7733–7738.

[6] H. Eldridge, M. De Donno, C. Champod, Testing the accuracy and reliability of palmar friction ridge comparisons – a black box study, Forensic Sci. Int. 318 (2021), 110457.

[7] K.L. Monson, E.D. Smith, E.M. Peters, Accuracy of comparison decisions by forensic firearms examiners, J. Forensic Sci. 68 (1) (2023) 86–100.

[8] R.A. Hicklin, L. Eisenhart, N. Richetelli, M.D. Miller, P. Belcastro, T.M. Burkes, C. L. Parks, M.A. Smith, J. Buscaglia, E.M. Peters, et al., Accuracy and reliability of forensic handwriting comparisons, Proc. Natl. Acad. Sci. USA 119 (32) (2022), e2119944119.

[9] G.H. Fischer, I.W. Molenaar, Rasch Models: Foundations, Recent Developments, and Applications, Springer Science & Business Media, New York, 2012.

[10] W.J. Van der Linden, R. Hambleton, Handbook of Item Response Theory, Taylor & Francis Group, 1997.

[11] A. Luby, Decision making in forensic identification tasks, in: S. Tyner, H. Hofmann (Eds.), Open Forensic Science in R, rOpenSci, US, 2019 (chapter 13).

[12] H. Hofmann, A. Carriquiry, S. Vanderplas, Treatment of inconclusives in the afte range of conclusions, Law Probab. Risk 19 (3–4) (2020) 317–364.

[13] I.E. Dror, N. Scurich, (mis) use of scientific measurements in forensic science, Forensic Sci. Int.: Synergy 2 (2020) 333–338.

[14] A.H. Dorfman, R. Valliant, Inconclusives, errors, and error rates in forensic firearms analysis: three statistical perspectives, Forensic Sci. Int.: Synergy (2022), 100273.

[15] E.J. Mattijssen, C.L. Witteman, C.E. Berger, N.W. Brand, R.D. Stoel, Validity and reliability of forensic firearm examiners, Forensic Sci. Int. 307 (2020), 110112.

[16] P. De Boeck, I. Partchev, Irtrees: tree-based item response models of the glmm family, J. Stat. Softw. Code Snippets 48 (1) (2012) 1–28.

[17] M. Jeon, P. De Boeck, A generalized item response tree model for psychological assessments, Behav. Res. Methods 48 (3) (2016) 1070–1085.

[18] A.S. Luby, J.B. Kadane, Proficiency testing of fingerprint examiners with Bayesian item response theory, Law Probab. Risk 17 (2) (2018) 111–121.

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

[20] T. Busey, M. Coon, Not All Identification Conclusions Are Equal: Quantifying the Strength of Fingerprint Decisions, Forensic Science International, 2023, 111543.

[21] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Repeatability and reproducibility of decisions by latent fingerprint examiners, PLoS One 7 (3) (2012), e32800.

[22] B.T. Ulery, R.A. Hicklin, M.A. Roberts, J. Buscaglia, Measuring what latent fingerprint examiners consider sufficient information for individualization determinations, PLoS One 9 (11) (2014), e110179.

[23] N. Kalka, M. Beachler, R. Hicklin, Lqmetric: a latent fingerprint quality metric for predicting afis performance and assessing the value of latent fingerprints, J. Forensic Ident. 70 (2021) 443–463.

[24] G. Rasch, Probabilistic Models for Some Intelligence and Attainment Tests, University of Chicago Press, Chicago, 1960.

[25] F.M. Lord, Applications of Item Response Theory to Practical Testing Problems, Routledge, 1980.

[26] P. de Boeck, M. Wilson, Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach, Springer, New York, 2004.

[27] I. Partchev, P. De Boeck, Can fast and slow intelligence be differentiated? Intelligence 40 (1) (2012) 23–32.

[28] H. Plieninger, T. Meiser, Validity of multiprocess irt models for separating content and response styles, Educ. Psychol. Meas. 74 (5) (2014) 875–899.

[29] M. Jeon, P. De Boeck, W. van der Linden, Modeling answer change behavior: an application of a generalized item response tree model, J. Educ. Behav. Stat. 42 (4) (2017) 467–490.

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

[31] A. Gelman, X.-L. Meng, H. Stern, Posterior predictive assessment of model fitness via realized discrepancies, Stat. Sin. (1996) 733–760.

[32] S. Sinharay, M.S. Johnson, H.S. Stern, Posterior predictive assessment of item response theory models, Appl. Psychol. Meas. 30 (4) (2006) 298–321.

[33] DOJ, United States Department of Justice Statement on the PCAST Report: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, United States Department of Justice, 2021. Technical report.

[34] AAAS, Forensic Science Assessments: A Quality and Gap Analysis - Latent Fingerprint Examination, 2017. Technical report, (prepared by William Thompson, John Black, Anil Jain, and Joseph Kadane).

[35] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2023.

[36] Posit team, RStudio: *Integrated Development Environment For R*. Posit Software, PBC, Boston, MA, 2023.

[37] H. Wickham, M. Averick, J. Bryan, W. Chang, L.D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T.L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D.P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, Welcome to the tidyverse, J. Open Source Softw. 4 (43) (2019) 1686.

[38] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2016.

[39] B. Schloerke, D. Cook, J. Larmarange, F. Briatte, M. Marbach, E. Thoen, A. Elberg, J. Crowley, *GGally: Extension To 'ggplot2'*. R Package version 2.1.2, 2021.

[40] T.L. Pedersen, Patchwork: the Composer of Plots, 2022. R package version 1.1.2.

[41] Stan Development Team, RStan: the R Interface to Stan. R Package Version 2.18.2, 2018.

[42] Stan Development Team, Stan Modeling Language Users Guide and Reference Manual, 2018.

[43] D. Lewandowski, D. Kurowicka, H. Joe, Generating random correlation matrices based on vines and extended onion method, J. Multivariate Anal. 100 (9) (2009) 1989–2001.