2019

Decision-Making In Forensic Identification Tasks

Amanda Luby
Swarthmore College, aluby1@swarthmore.edu

Chapter 8 Decision-making in Forensic Identification Tasks

sctyner.github.io/OpenForSciR/decision-making.html

Amanda Luby

8.1 Introduction

Although forensic measurement and analysis tools are increasingly accurate and objective, many final decisions are largely left to individual examiners (PCAST 2016). Human decision-makers will continue to play a central role in forensic science for the foreseeable future, and it is unrealistic to assume that, within the United States' current criminal justice system,

- there are no differences in the decision-making process between examiners,
- day-to-day forensic decision-making tasks are equally difficult, or
- human decision-making can be removed from the process entirely.

The role of human decisions in forensic science is perhaps most studied in the fingerprint domain, which will be the focus of this chapter. High-profile examples of misidentification have inspired studies showing that fingerprint examiners, like all humans, may be

susceptible to biased instructions and unreliable in final decisions (Dror and Rosenthal 2008) or influenced by external factors or contextual information (Dror, Charlton, and Péron 2006; Dror and Cole 2010). These studies contradict common perceptions of the accuracy of fingerprint examination, and demonstrate that fingerprint analysis is far from error-free.

Although fingerprint examination is the focus of this chapter, it is not the only forensic domain that relies on human decision-making. Firearms examination (see, e.g., NRC (2009b) pg. 150-155) is similar to latent print examination in many ways, particularly in that examiners rely on pattern evidence to determine whether two cartridges originated from the same source. Handwriting comparison (see National Research Council (2009b) pg. 163-167 on "Questioned Document Examination" and Stoel et al. (2010) for discussion) consists of examiners determining whether two samples of handwriting were authored by the same person, taking potential forgery or disguise into account. A third example is interpreting mixtures of DNA evidence (see PCAST (2016) Section 5.2). A DNA mixture is a biological sample that contains DNA from two or more donors and requires analysts to make subjective decisions to determine how many individuals contributed to the DNA profile. Due to these currently unavoidable human factors, the President's Council of Advisors on Science and Technology (2016) recommended increased "black box" error rate studies for these and other subjective forensic science methods.

The FBI "Black Box" study (Bradford T. Ulery et al. 2011) was the first large-scale study performed to assess the accuracy and reliability of latent print examiners' decisions. The questions included a range of attributes and quality seen in casework, and were representative of searches from an automated fingerprint identification system. The overall false positive rate in the study was 0.1% and the overall false negative rate was 7.5%. These computed quantities, however, have excluded all "inconclusive" responses (i.e. neither identifications nor exclusions). This is noteworthy, as nearly a third of all responses were inconclusive and respondents varied on how often they reported inconclusives. Respondents who report a large number of inconclusives, and only make identification or exclusion decisions for the most pristine prints, will likely make far fewer false positive and false negative decisions than respondents who reported fewer inconclusives. The authors of the study also note that it is difficult to compare the error rates and inconclusive rates of individual examiners because each examiner saw a different set of fingerprint images (see Appendix 3 of Bradford T. Ulery et al. (2011)). In other words, it would be unfair to compare the error rate of someone who was given a set of "easy" questions to the error rate of someone who was given a set of "difficult" questions. A better measure of examiner skill would account for both error rates and difficulty of prints that were examined.

Accurately measuring proficiency, or examiner skill, is valuable not only for determining whether a forensic examiner has met baseline competency requirements, but for training purposes as well. Personalized feedback after participating in a study could lead to targeted

training for examiners in order to improve their proficiency. Additionally, if proficiency is not accounted for among a group of study participants, which often include trainees or non-experts as well as experienced examiners, the overall results from the study may be biased.

There also exist substantial differences in the difficulty of forensic evaluation tasks. Properties of the evidence, such as the quality, quantity, concentration, or rarity of characteristics may make it easier or harder to evaluate. Some evidence, regardless of how skilled the examiner is, will not have enough information to result in an identification or exclusion in a comparison task. An inconclusive response, in this case, should be treated as the "correct" response. Inconclusive responses on more straightforward identification tasks, on the other hand, may be treated as mistakes.

Methods for analyzing forensic decision-making data should thus provide estimates for both participant proficiency and evidence difficulty. *Item response models*, a class of statistical methods used prominently in educational testing, have been proposed for use in forensic science for these reasons (Kerkhoff et al. 2015). Luby and Kadane (2018) provided the first item response analysis for forensic proficiency test data, and we improve and extend upon that work by - analyzing a different fingerprint identification study that includes richer data on decision-making, and - extending the range of models considered.

The remainder of the chapter is organized as follows: Section 8.1.1 gives a brief overview of Item Response Models, Section 8.2 provides an overview on how decision-making data is collected in forensic science, and Section 8.3 describes an R package that can be used to fit these models. Section 8.4 describes how conclusions are drawn from an Item Response analysis, and Section 8.5 gives an example IRT analysis of the FBI "Black Box" study.

8.1.1 A Brief Overview of Item Response Models

For PP individuals responding to II test items, we can express the binary responses (i.e. correct/incorrect) as a P×IP×I matrix, YY. Item Response Theory (IRT) is based on the idea that the probability of a correct response depends on individual *proficiency*, θp,p=1,…,Pθp,p=1,…,P, and item *difficulty*, bi,i=1,…Ibi,i=1,…I.

8.1.1.1 Rasch Model

The Rasch Model (Rasch 1960; Fischer and Molenaar 2012) is a relatively simple yet powerful item response model, and serves as the basis for extensions introduced later. The probability of a correct response is modeled as a logistic function of the difference between the participant proficiency, θpθp (p=1,…,Pp=1,…,P), and the item difficulty, bibi (i=1,…,Ii=1,…,I):

P(Ypi=1)=11−exp(−(θp−bi)).(8.1)

P(Ypi=1)=11−exp(−(θp−bi)).

To identify the model, we shall use the convention of constraining the mean of the participant parameters (μ_θ) to be equal to zero. This allows for a nice interpretation of both participant and item parameters relative to the "average participant". If $\theta_p > 0$, participant p is of "above average" proficiency and if $\theta_p < 0$, participant p is of "below average" proficiency. Similarly, if $b_i < 0$ question i is an "easier" question and the average participant is more likely to correctly answer question i. If $b_i > 0$ then question i is a more "difficult" question and the average participant is less likely to correctly answer question i. Other common conventions for identifying the model include setting a particular b_i or the mean of the b_is equal to zero.

The item characteristic curve (ICC) describes the relationship between proficiency and performance on a particular item (see Figure 8.1 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.

Note that Equation (8.1) also describes a generalized linear model (GLM), where $\theta_p - b_i$ is the linear component, with a logit link function. By formulating the Rasch Model as a hierarchical GLM with prior distributions on both θ_p and b_i, the identifiability problem is solved. We assign $\theta_p \sim N(0, \sigma^2_\theta)$ and $b_i \sim N(\mu_b, \sigma^2_b)$, although more complicated prior distributions are certainly possible.

The *two-parameter logistic model* (2PL) and *three-parameter logistic model* (3PL) are additional popular item response models (Lord 1980). 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. We omit a full discussion of these models here, but further reading may be found in van der Linden and Hambleton (2013) and Boeck and Wilson (2004).

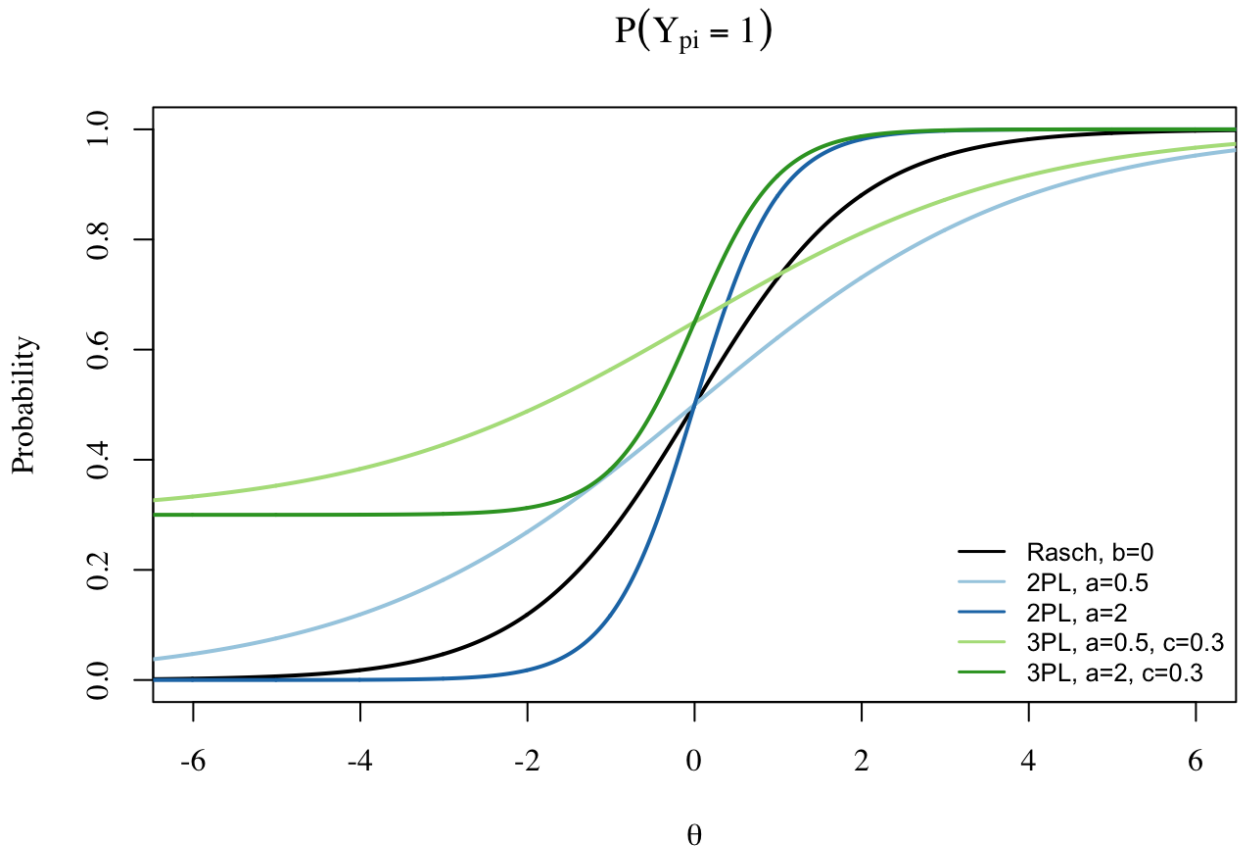


Figure 8.1: Item Characteristic Curve (ICC) examples for the Rasch, 2PL, and 3PL models.

8.1.1.2 Partial Credit Model

The *partial credit model* (PCM) (Masters 1982) is distinct from the models discussed above because it allows for the response variable, YpiYpi, to take additional values beyond zero (incorrect) and one (correct). This is especially useful for modeling partially correct responses, although may be applied in other contexts where the responses can be ordered. When YpiYpi is binary, the partial credit model is equivalent to the Rasch model. Under the PCM, the probability of response YpiYpi depends on θpθp, the proficiency of participant pp as in the above models; mimi, the maximum score for item ii (and the number of step parameters); and βilβil, the lthlth step parameter for item ii (l=0,…,mil=0,…,mi):

P(Ypi=0)=11+∑mik=1exp∑kl=1(θp−βil)
P(Ypi=0)=11+∑k=1miexp∑l=1k(θp−βil)

P(Ypi=y,y>0)=exp∑yl=1(θp−βil)1+∑mik=1exp∑kl=1(θp−βil).(8.2)
P(Ypi=y,y>0)=exp∑l=1y(θp−βil)1+∑k=1miexp∑l=1k(θp−βil).

An example PCM is shown in Figure 8.2 by plotting the probabilities of observing each of three categories as a function of θpθp (analogous to the ICC curves above).

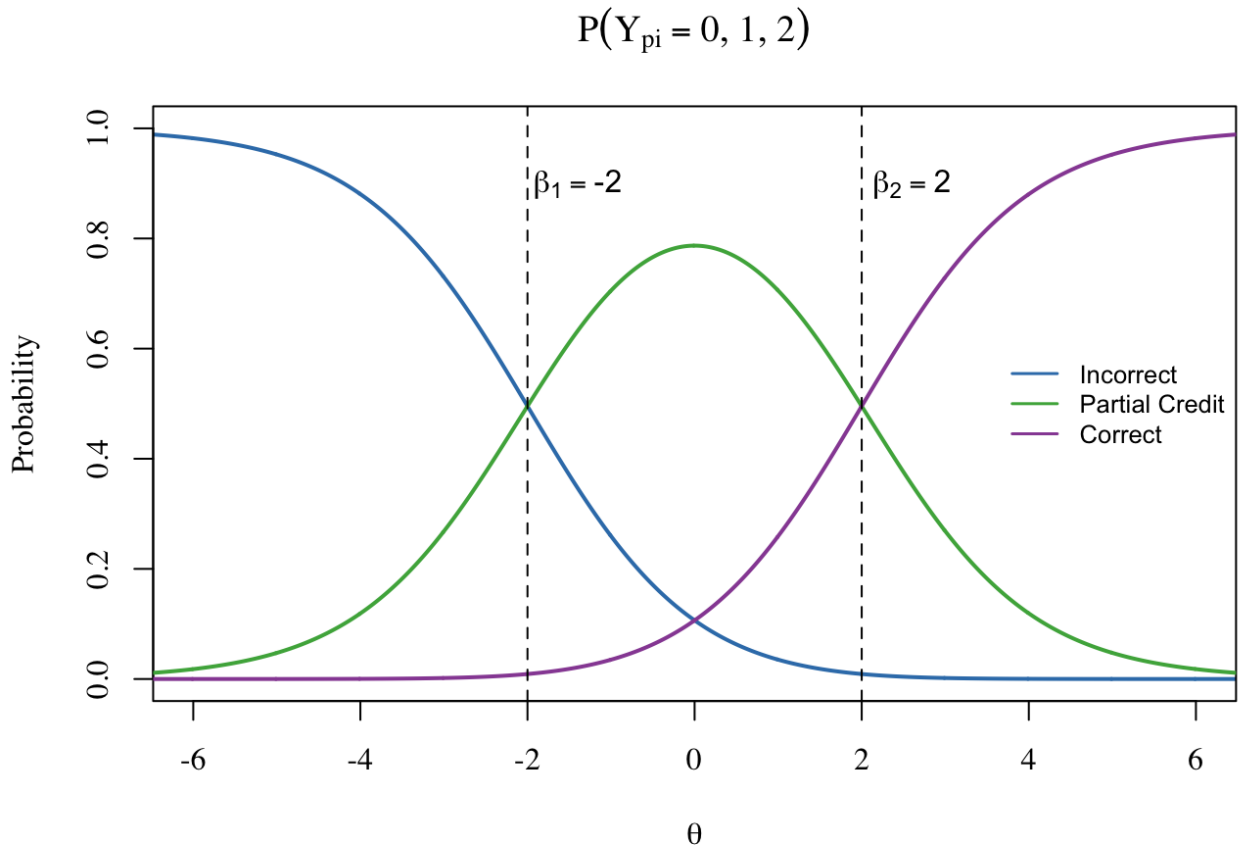


Figure 8.2: Category response functions for the PCM.

8.2 Data

The vast majority of forensic decision-making occurs in casework, information about which is not often made available to researchers due to privacy concerns. Outside of casework, data on forensic science decision-making is collected through proficiency test results and in error rate studies. *Proficiency tests* are periodic competency exams that must be completed for forensic laboratories to maintain their accreditation. *Error rate studies* are independent research studies designed to measure casework error rates. As their names suggest, these two data collection scenarios serve completely different purposes. Proficiency tests are designed to assess *basic competency* of individuals, and mistakes are rare. Error rate studies are designed to mimic the difficulty of evidence in casework and estimate the *overall error rate*, aggregating over many individuals, and mistakes are more common by design.

Proficiency exams consist of a large number of participants (often >400>400) responding to a small set of questions (often <20<20). Since every participant answers every question, we can assess participant proficiency and question difficulty using the observed scores. As proficiency exams are designed to assess basic competency, most questions are relatively easy and the vast majority of participants score 100%. Error rate studies, on the other hand, consist of a smaller number of participants (fewer than 200200) and a larger pool of questions (more than 500500). The questions are designed to be difficult, and every participant does not answer every question, which makes determining participant proficiency and question difficulty a more complicated task.

Results from both proficiency tests and error rate studies can be represented as a set of individuals responding to several items, in which responses can be scored as correct or incorrect. This is not unlike an educational testing scenario where students (individuals) answer questions (items) either correctly or incorrectly. There is a rich body of statistical methods for estimating student proficiency and item difficulty from test responses. Item Response Theory (IRT) is used extensively in educational testing to study the relationship between an individual's (unobserved) proficiency and their performance on varying tasks. IRT is an especially useful tool to estimate participant proficiencies and question difficulties when participants do not necessarily answer the same set of questions.

8.3 R Packages

The case study makes use of the `blackboxstudyR` R package (Luby 2019), which provides functions for working with the FBI black box data, implementations of basic IRT models in Stan (Guo, Gabry, and Goodrich 2018), and plotting functions for results.

The primary functions of `blackboxstudyR` include:

- `score_bb_data()`: Scores the FBI "Black Box" data under one of five scoring schemes.
- `irt_data_bb()`: Formats the FBI "Black Box" data into a form appropriate for fitting a Stan model.
- `fit_irt()`: Wrapper for Stan to fit standard IRT models to data (does not be the FBI data). Models currently available are:
    - Rasch Model (Section 8.1.1.1)
    - 2PL Model (Section 8.1.1.1)
    - Partial Credit Model (Section 8.1.1.2)
- `plot_difficulty_posteriors` and `plot_proficiency_posteriors`: Plot posterior intervals for difficulty and proficiency estimates, respectively.

8.4 Drawing Conclusions

An IRT analysis produces estimates of both participant proficiency and item difficulty. As mentioned previously, this property is especially useful for settings where participants respond to different subsets of items, as it allows all participants to be compared on the same scale.

By comparing the estimated proficiency to more traditional measures of participant performance (e.g. false positive rate or false negative rate), we can see whether there are aspects captured by proficiency that are not captured in other measures. For instance, the false positive rate and the false negative rate contain no information about the inconclusive rate, while IRT does implicitly, as it accounts for the number of question answered by each participant.

In the forensic science setting, completing an IRT analysis will often include an additional step of choosing how the data should be scored. For example, should inconclusive responses be scored as incorrect or treated as missing? An additional question we may wish to answer is, “Which scoring scheme is most appropriate for the setting at hand?” In some cases, the optimal scoring scheme may be determined using expert knowledge, or by specifying the expected answers to each item beforehand. In other cases, it may not be possible to determine the optimal scoring scheme before fitting an IRT model. In those cases, multiple scoring methods should be used to fit an IRT model, and the results from each model should be compared and contrasted.

8.5 Case Study

We use the FBI “black box” data (`blackboxstudyR::TestResponses`) for our case study. `TestResponses` is a data frame in which each row corresponds to an examiner, each column represents the item, and the value in each unique combination of row and column is the examiner’s response to that item. In addition to the examiner ID (`Examiner_ID`) and item ID (`Pair_ID`), the data contains:

- `Mating`: whether the pair of prints were “Mates” (same source) or “Non-mates” (different source)
- `Latent_Value`: the examiner’s assessment of the value of the print (NV = No Value, VEO = Value for Exclusion Only, VID = Value for Individualization)
- `Compare_Value`: the examiner’s assessment of whether the pair of prints is an “Exclusion”, “Inconclusive” or “Individualization”

- `Inconclusive_Reason`: If inconclusive, the reason for the inconclusive
  - "Close": *The correspondence of features is supportive of the conclusion that the two impressions originated from the same source, but not the extent sufficient for individualization.*
  - "Insufficient": *Potentially corresponding areas are present, but there is insufficient information present.* Participants were told to select this reason if the reference print was not of value.
  - "No Overlap": *No overlapping area between the latent and reference*
- `Exclusion_Reason`: If exclusion, the reason for the exclusion
  - "Minutiae"
  - "Pattern"
- `Difficulty`: Reported difficulty ranging from "A_Obvious" to "E_VeryDifficult"

In order to fit an IRT model, we must first score the data. Responses are scored as correct if they are true identifications (`Mating == Mates` and `Compare_Value == Individualization`) or exclusions (`Mating == Non-mates` and `Compare_Value == Exclusion`). Similarly, responses are scored as incorrect if they are false identifications (`Mating == Non-mates` and `Compare_Value == Individualization`) or false exclusions (`Mating == Mates` and `Compare_Value == Exclusion`).

Inconclusive responses, which are never keyed as correct responses, complicate the scoring of the exam due to both their ambiguity and prevalence. There are a large number of inconclusive answers (4907 of 17121 responses), and examiners vary on which latent print pairs are inconclusive.

The `blackboxstudyR` package includes five methods to score inconclusive responses:

1. Score all inconclusive responses as incorrect (`inconclusive_incorrect`). This may penalize participants who were shown more vague or harder questions and therefore reported more inconclusives.
2. Treat inconclusive responses as missing completely at random (`inconclusive_mcar`). This decreases the amount of data included in the analysis, and does not explicitly penalize examiners who report many inconclusives. This is the scoring method most similar to the method used in Bradford T. Ulery et al. (2011) to compute false positive and false negative rates.
3. Score inconclusive as correct if the reason given for an inconclusive is "correct". Since the ground truth "correct" inconclusive reason is unknown, the consensus reason from other inconclusive responses for that question is used. If no consensus reason exists, the inconclusive response was scored in one of two ways:
   1. Treat inconclusive responses as incorrect if no consensus reason exists (`no_consensus_incorrect`).
   2. Treat inconclusive responses as missing completely at random if no consensus reason exists (`no_consensus_mcar`).

4. Score inconclusive responses as “partial credit” (`partial_credit`).

In the remainder of the case study we will 1. demonstrate how to fit an IRT model in R, 2. illustrate how IRT analysis complements an error rate analysis by accounting for participants seeing different sets of questions, and 3. show how different scoring methods can change results from an IRT analysis.

8.5.1 Fitting the IRT model

We’ll proceed with an IRT analysis of the data under the `inconclusive_mcar` scoring scheme, which is analogous to how the data were scored under Bradford T. Ulery et al. (2011).

```
im_scored <- score_bb_data(TestResponses, "inconclusive_mcar")
```

Scoring the black box data as above gives us the response variable (yy). The `irt_data_bb` function takes the original black box data, along with the scored variable produced by `score_bb_data`, and produces a list object in the form needed by Stan to fit the IRT models. If you wish to fit the models on a different set of data, you can do so if the dataset has been formatted as a list object with the same attributes as the `irt_data_bb` function output (see package documentation for additional details).

```
im_data <- irt_data_bb(TestResponses, im_scored)
```

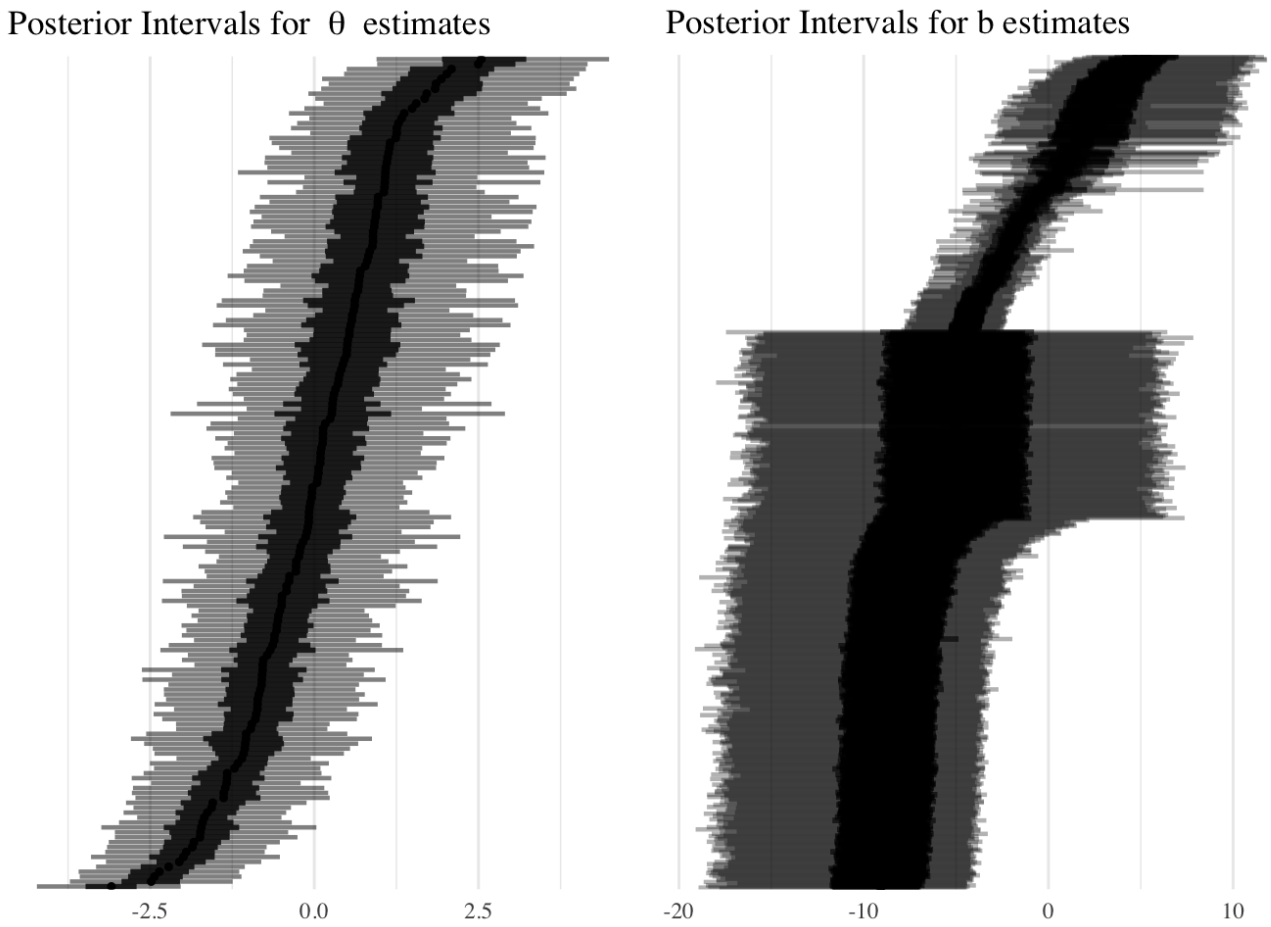
We can now use `fit_irt` to fit the Rasch models.

```
im_model <- fit_rasch(im_data, iterations = 600, n_chains = 4)
```

In practice, it is necessary to ensure that the MCMC sampler has converged using a variety of diagnostics. We omit these steps here for brevity, but the `blackboxstudyR` package will include a vignette detailing this process, or see e.g. Gelman et al. (2013).

After the model has been fit, we can plot the posterior distributions of difficulties and proficiencies:

```
p1 <- plot_proficiency_posteriors(im_samples) + my_theme
p2 <- plot_difficulty_posteriors(im_samples) + my_theme
ggarrange(p1, p2, ncol = 2)
```



The lighter gray interval represents the 95% posterior interval and the black interval represents the 50% posterior interval. If we examine the posterior intervals for the difficulty estimates (bb), we can see groups which have noticeably larger intervals, and thus more uncertainty regarding the estimate: 1. those on the bottom left 2. those on the upper right, and 3. those in the middle.

These three groups of uncertain estimates correspond to: 1. the questions that every participant answered correctly, 2. the questions that every participant answered incorrectly, and 3. the questions that every participant reported as an “inconclusive” or “no value”.

8.5.2 IRT complements an error rate analysis

The original analysis of the FBI “Black Box” Study (see Bradford T. Ulery et al. 2011) did not include analysis of participant error rates, because each participant saw a different question set. Since proficiency accounts for the difficulty of question sets, however, we can directly compare participant proficiencies to each other, and also see how error rates and proficiency are related.

First, we compute the observed person scores.

```
obs_p_score <- bb_person_score(TestResponses, im_scored)
```

In order to use the `error_rate_analysis` function, we need to extract the median question difficulties from MCMC results.

```
q_diff <- apply(im_samples, 3, median)[grep("b\\[", names(apply(im_samples,
    3, median)))]
ex_error_rates <- error_rate_analysis(TestResponses, q_diff)
```

Now, we can plot the proficiency estimates (with 95% posterior intervals) against the results from a traditional error rate analysis.

```
p1 <- person_mcmc_intervals(im_samples) %>% right_join(., obs_p_score, by = "exID")
%>%
    full_join(., ex_error_rates, by = "exID") %>% dplyr::select(., score, m,
    ll, hh, exID, avg_diff, fpr, fnr) %>% ggplot(., aes(x = fpr, y = m, ymin = ll,
    ymax = hh)) + geom_pointrange(size = 0.3) + labs(x = "False Positive Rate",
    y = "Proficiency Estimate") + my_theme

p2 <- person_mcmc_intervals(im_samples) %>% right_join(., obs_p_score, by = "exID")
%>%
    full_join(., ex_error_rates, by = "exID") %>% dplyr::select(., score, m,
    ll, hh, exID, avg_diff, fpr, fnr) %>% ggplot(., aes(x = fnr, y = m, ymin = ll,
    ymax = hh, color = fpr > 0)) + geom_pointrange(size = 0.3) + labs(x = "False
Negative Rate",
    y = "Proficiency Estimate") + scale_colour_manual(values = c("black",
"steelblue")) +
    my_theme + theme(legend.position = "none")

ggarrange(p1, p2, ncol = 2)
```

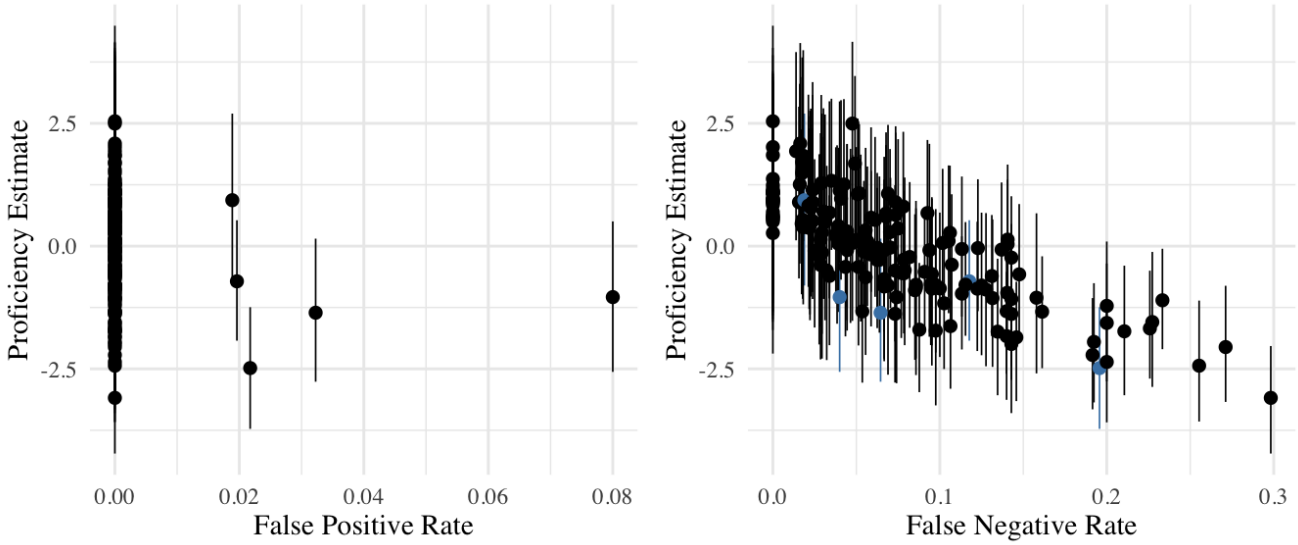


Figure 8.3: Proficiency vs False Positive Rate (left) and False Negative Rate (right)

Figure 8.3 shows proficiency against the false positive rate (left) and false negative rate (right). Those participants who made at least one false positive error are colored in blue on the right side plot. We see that one of the participants who made a false positive error still received a relatively large proficiency estimate due to having such a small false negative rate.

If, instead of looking error rates for each participant, we examine observed scores, the estimated proficiencies correlate with the observed score (Figure 8.4. That is, participants with a higher observed score are generally given larger proficiency estimates than participants with lower scores. There are, however, cases where participants scored roughly the same on the study but are given vastly different proficiency estimates. For example, the highlighted participants in the right plot above all scored between 94% and 96%, but their estimated proficiencies range from −1.25−1.25 to 2.52.5.

```
p1 <- person_mcmc_intervals(im_samples) %>% right_join(obs_p_score, by = "exID") %>%
    ggplot(aes(x = score, y = m, ymin = ll, ymax = hh)) + geom_pointrange(size = 0.3)
+
    labs(x = "Observed Score", y = "Proficiency Estimate") + my_theme

p2 <- person_mcmc_intervals(im_samples) %>% right_join(obs_p_score, by = "exID") %>%
    ggplot(aes(x = score, y = m, ymin = ll, ymax = hh)) + geom_pointrange(size = 0.3)
+
    gghighlight(score < 0.96 & score > 0.94) + labs(x = "Observed Score", y =
"Proficiency Estimate") +
    my_theme

ggarrange(p1, p2, ncol = 2)
```

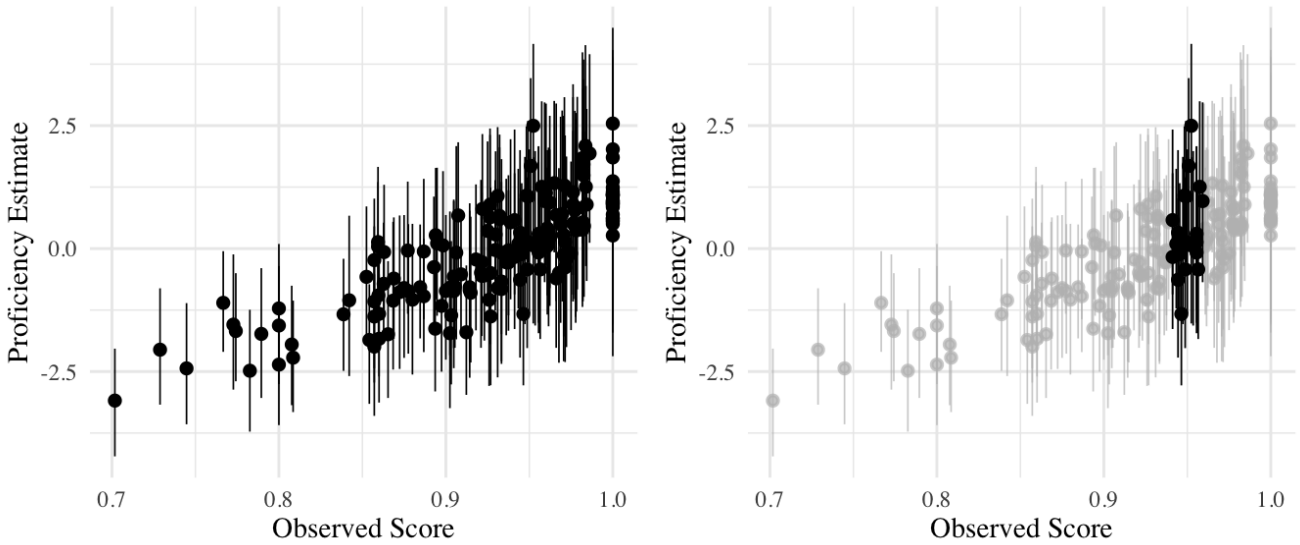


Figure 8.4: Proficiency vs Observed Score

If we examine those participants who scored between 94% and 96% more closely, we can see that the discrepancies in their proficiencies are largely explained by the difficulty of the specific question set they saw. This is evidenced by the positive trend in Figure 8.5. In addition to the observed score and difficulty of the question set, the number of questions the participant answers conclusively (i.e. individualization or exclusion) also plays a role in the proficiency estimate. Participants who are conclusive more often generally receive higher estimates of proficiency than participants who are conclusive less often.

```
person_mcmc_intervals(im_samples) %>% right_join(obs_p_score, by = "exID") %>%
    full_join(ex_error_rates, by = "exID") %>% dplyr::select(score, m, ll, hh,
    exID, avg_diff, pct_skipped) %>% filter(score < 0.96 & score > 0.94) %>%
    ggplot(aes(x = avg_diff, y = m, ymin = ll, ymax = hh, col = 1 - pct_skipped)) +
    geom_pointrange(size = 0.3) + labs(x = "Avg Q Difficulty", y = "Proficiency
Estimate",
    color = "% Conclusive") + my_theme
```

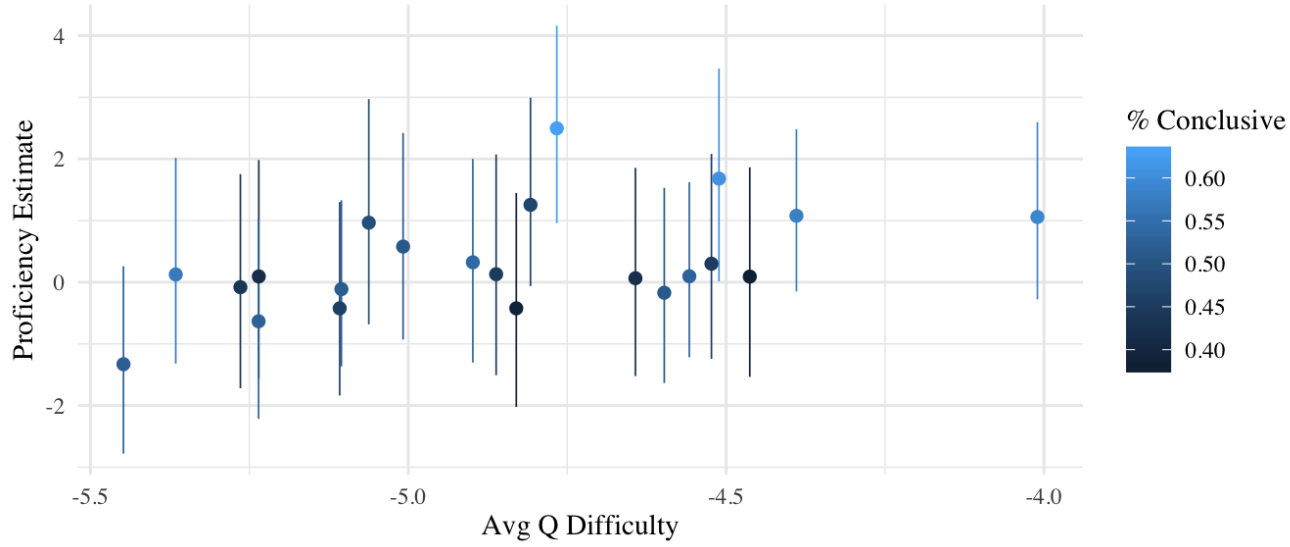


Figure 8.5: Proficiency vs Average Question Difficulty, for participants with observed score between 94 and 96 percent correct.

8.5.3 Scoring method affects proficiency estimates

To illustrate the difference in results between different scoring methods, we'll now score the data and fit models in two more ways: `no_consensus_incorrect` and `partial_credit`.

```
nci_scored <- score_bb_data(TestResponses, "no_consensus_incorrect")
nci_data <- irt_data_bb(TestResponses, nci_scored)
pc_scored <- score_bb_data(TestResponses, "partial_credit")
pc_data <- irt_data_bb(TestResponses, pc_scored)
```

We use `fit_rasch` to fit the Rasch model to the `no_consensus_incorrect` data, and since the `partial_credit` data has three outcomes (correct, inconclusive, or incorrect) instead of only two (correct/incorrect), we use `fit_pcm` to fit a partial credit model to the data.

```
nci_model <- fit_rasch(nci_data, iterations = 1000, n_chains = 4)
pc_model <- fit_pcm(pc_data, iterations = 1000, n_chains = 4)
```

We can examine the proficiency estimates and observed scores for each participant under each of the three scoring schemes, similar to Figure 8.4 above. Under the partial credit scoring scheme, a correct identification/exclusion is scored as a “2”, an inconclusive response is scored as a “1” and an incorrect identification/exclusion is scored as a “0”. The observed score is then computed by (#Correct+#Inconclusive)/(2×#Responses) (#Correct+#Inconclusive)/(2×#Responses) to scale the score to be between 0 and 1.

```
p_score_im <- bb_person_score(TestResponses, im_scored)
p_score_im <- person_mcmc_intervals(blackboxstudyR::im_samples) %>%
right_join(p_score_im,
    by = "exID") %>% mutate(scoring = rep("im", nrow(p_score_im)))

p_score_nci <- bb_person_score(TestResponses, nci_scored)
p_score_nci <- person_mcmc_intervals(blackboxstudyR::nci_samples) %>%
right_join(p_score_nci,
    by = "exID") %>% mutate(scoring = rep("nci", nrow(p_score_nci)))

p_score_pc <- bb_person_score(TestResponses, pc_scored)
p_score_pc <- person_mcmc_intervals(blackboxstudyR::pc_samples) %>%
right_join(p_score_pc,
    by = "exID") %>% mutate(scoring = rep("pc", nrow(p_score_pc)))

p1 <- p_score_im %>% bind_rows(p_score_nci) %>% bind_rows(p_score_pc) %>%
ggplot(aes(x = score,
    y = m, ymin = ll, ymax = hh, col = scoring)) + geom_pointrange(size = 0.3,
    alpha = 0.5) + labs(x = "Observed Score", y = "Estimated Proficiency") +
    my_theme

p2 <- p_score_im %>% bind_rows(p_score_nci) %>% bind_rows(p_score_pc) %>%
group_by(exID) %>%
    ggplot(aes(x = score, y = m, ymin = ll, ymax = hh, col = scoring, group = exID))
+
    geom_pointrange() + gghighlight(hh < -0.5, use_group_by = FALSE) + geom_line(col
= "black",
    linetype = "dotted") + labs(x = "Observed Score", y = "Estimated Proficiency") +
    geom_hline(yintercept = -0.5, col = "darkred", linetype = "dashed") + my_theme

ggarrange(p1, p2, ncol = 2, common.legend = TRUE, legend = "bottom")
```

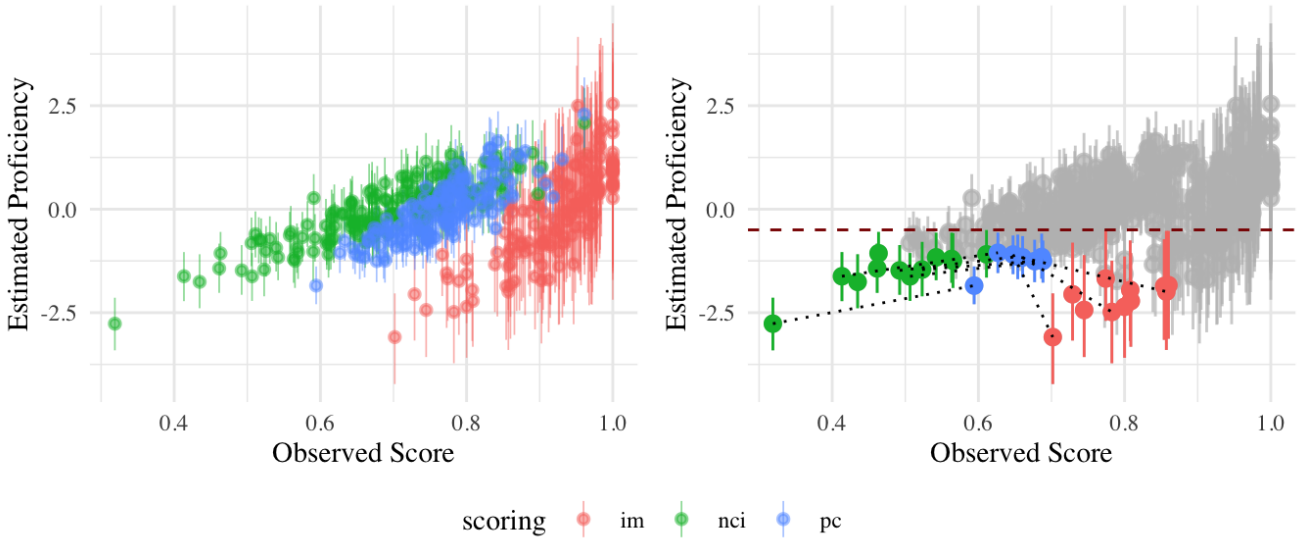


Figure 8.6: Proficiency vs Observed Score for each of three scoring schemes

Treating the inconclusives as missing (“im”), leads to both the smallest range of observed scores and largest range of estimated proficiencies. Harsher scoring methods (e.g. `no consensus incorrect` (“nci”)) do not necessarily lead to lower estimated proficiencies. For instance, the participants who scored around 45% under the “nci” scoring method (in green) are given higher proficiency estimates than the participant who scored 70% under the “im” scoring method. The scoring method thus affects the proficiency estimates in a somewhat non-intuitive way, as larger ranges of observed scores do not necessarily correspond to larger ranges of proficiency estimates.

Also note that the uncertainty intervals under the “im” scoring scheme are noticeably larger than under the other scoring schemes. This is because the `inconclusive_mcar` scheme treats all of the inconclusives, nearly a third of the data, as missing. This missingness is completely uninformative when estimating the difficulty and proficiency estimates. Under the other scoring schemes (`no consensus incorrect` and `partial credit`) the inconclusive responses are never treated as missing, leading to a larger number of observations per participant and therefore a smaller amount of uncertainty in the proficiency estimate.

The range of proficiencies under different scoring schemes and the uncertainty intervals for the proficiency estimates both have substantial implications if we consider setting a “mastery level” for participants. As an example, let’s consider setting the mastery threshold at −0.5−0.5. We will then say examiners have not demonstrated mastery if the upper end of their proficiency uncertainty estimate is below −0.5−0.5, illustrated in the right plot of Figure 8.6.

The number of examiners that have not demonstrated mastery varies based on the scoring method used (11 for “nci”, 8 for “pc” and 11 for “im”) due to the variation in range of proficiency estimates. Additionally, for each of the scoring schemes, there are a number of

examiners that did achieve mastery with the same observed score as those that did not demonstrate mastery. This is due to a main feature of item response models discussed earlier: participants that answered more difficult questions are given higher proficiency estimates than participants that answered the same number of easier questions.

We've also drawn dotted lines between proficiency estimates that correspond to the same person. Note that many of the participants who do not achieve mastery under one scoring scheme *do* achieve mastery under the other scoring schemes, since not all of the points are connected by dotted lines. There are also a few participants who do not achieve mastery under any of the scoring schemes. This raises the question of how much the proficiency estimates change for each participant under the different scoring schemes.

The plot on the left in Figure 8.7 shows both a change in examiner proficiencies across scoring schemes (the lines connecting the proficiencies are not horizontal) as well as a change in the ordering of examiner proficiencies (the lines cross one another). That is, different scoring schemes affect examiner proficiencies in different ways.

The plot on the right illustrates participants that see substantial differences in their proficiency estimates under different scoring schemes. Examiners 105 and 3 benefit from the leniency in scoring when inconclusives are treated as missing ("im"). When inconclusives are scored as incorrect ("nci") or partial credit ("pc"), they see a substantial decrease in their proficiency due to reporting a high number of inconclusives and differing from other examiners in their reasoning for reporting inconclusives. Examiners 142, 60 and 110, on the other hand, are hurt by the leniency in scoring when inconclusives are treated as missing ("im"). Their proficiency estimates increase when inconclusives are scored as correct when they match the consensus reason ("nci") or are worth partial credit ("pc").

```
p1 <- p_score_im %>% bind_rows(p_score_nci) %>% bind_rows(p_score_pc) %>%
arrange(parameter) %>%
    mutate(id = rep(1:169, each = 3)) %>% dplyr::select(id, m, scoring) %>%
    spread(scoring, m) %>% mutate(max.diff = apply(cbind(abs(im - nci), abs(nci -
    pc), abs(im - pc)), 1, max)) %>% gather("model", "median", -c(id, max.diff)) %>%
    ggplot(aes(x = model, y = median, group = id, col = id)) + geom_point() +
    geom_line() + labs(x = "Scoring Method", y = "Estimated Proficiency") +
    my_theme

p2 <- p_score_im %>% bind_rows(p_score_nci) %>% bind_rows(p_score_pc) %>%
arrange(parameter) %>%
    mutate(id = rep(1:169, each = 3)) %>% dplyr::select(id, m, scoring) %>%
    spread(scoring, m) %>% mutate(max.diff = apply(cbind(abs(im - nci), abs(nci -
    pc), abs(im - pc)), 1, max)) %>% gather("model", "median", -c(id, max.diff)) %>%
    ggplot(aes(x = model, y = median, group = id, col = id)) + geom_point() +
    geom_line() + labs(x = "Scoring Method", y = "Estimated Proficiency") +
    gghighlight(max.diff > 1.95, label_key = id, use_group_by = FALSE) + my_theme

ggarrange(p1, p2, ncol = 2, common.legend = TRUE, legend = "bottom")
```

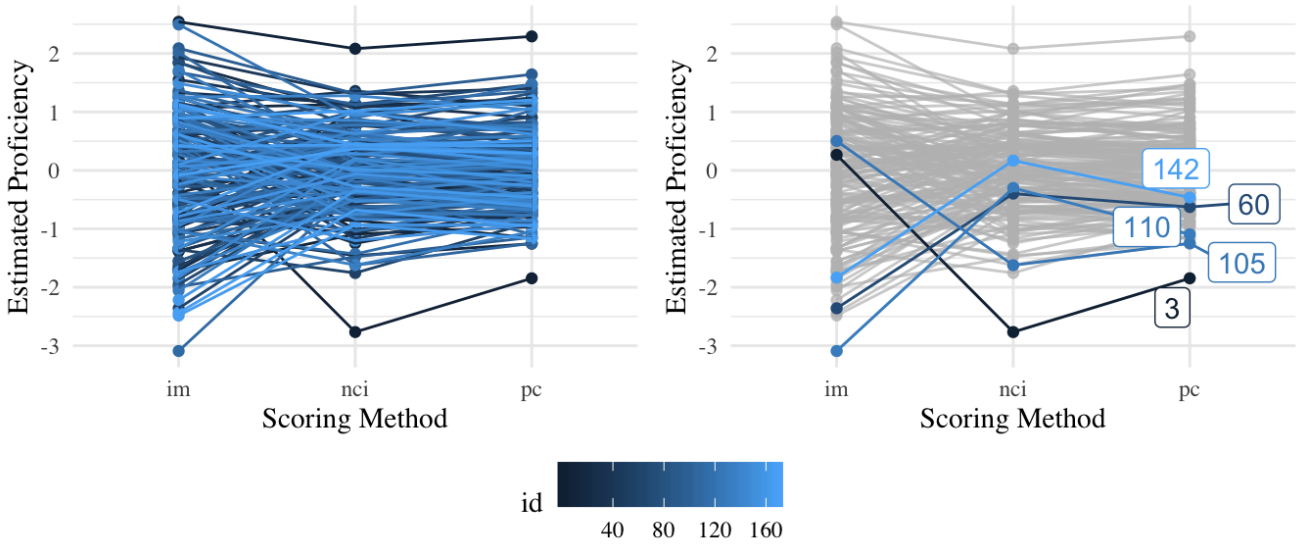


Figure 8.7: Change in proficiency for each examiner under the three scoring schemes. The right side plot has highlighted five examiners whose proficiency estimates change the most across schemes.

8.5.4 Discussion

We have provided an overview of human decision-making in forensic analyses, through the lens of latent print comparisons and the FBI “black box” study (Bradford T. Ulery et al. 2011). A brief overview of Item Response Theory (IRT), a class of models used extensively in educational testing, was introduced in Section 8.1.1. A case study is provided of an IRT analysis on the FBI black box’’ study in 8.5.

Results from an IRT analysis are largely consistent with conclusions from an error rate analysis. However, IRT provides substantially more information than a more traditional analysis, specifically through accounting for the difficulty of questions seen. Additionally, IRT implicitly accounts for the inconclusive rate of different participants and provides estimates of uncertainty for both participant proficiency and item difficulty. If IRT were to be adopted on a large scale, participants would be able to be directly compared even if they took different exams (for instance, proficiency exams in different years).

Three scoring schemes were presented in the case study, each of which leads to substantially different proficiency estimates across participants. Although IRT is a powerful tool for better understanding examiner performance on forensic identification tasks, we must be careful when choosing a scoring scheme. This is especially important for analyzing ambiguous responses, such as the inconclusive responses in the “black box” study.

References

PCAST, President's Council of Advisors on Science and Technology. 2016. "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods." Executive Office of The President's Council of Advisors on Science; Technology, Washington DC.

Dror, Itiel E, and Robert Rosenthal. 2008. "Meta-Analytically Quantifying the Reliability and Biasability of Forensic Experts." *Journal of Forensic Sciences* 53 (4): 900–903.

Dror, Itiel E, David Charlton, and Ailsa E Péron. 2006. "Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications." *Forensic Science International* 156 (1): 74–78.

Dror, Itiel E, and Simon A Cole. 2010. "The Vision in 'Blind' Justice: Expert Perception, Judgment, and Visual Cognition in Forensic Pattern Recognition." *Psychonomic Bulletin & Review* 17 (2): 161–67.

National Research Council. 2009b. *Strengthening Forensic Science in the United States: A Path Forward*. Washington, D.C.: The National Academies Press.

Stoel, Reinoud, Charles Berger, Elisa van den Heuvel, and Wil Fagel. 2010. "The Shaky Criticism of Forensic Handwriting Analysis."

Ulery, Bradford T., R. Austin Hicklin, JoAnn Buscaglia, and Maria Antonia Roberts. 2011. "Accuracy and Reliability of Forensic Latent Fingerprint Decisions." *Proceedings of the National Academy of Sciences* 108 (19): 7733–8.

Kerkhoff, W., R.D. Stoel, C.E.H. Berger, E.J.A.T. Mattijssen, R. Hermsen, N. Smits, and H.J.J. Hardy. 2015. "Design and Results of an Exploratory Double Blind Testing Program in Firearms Examination." *Science & Justice* 55 (6): 514–19. https://doi.org/https://doi.org/10.1016/j.scijus.2015.06.007.

Luby, Amanda S, and Joseph B Kadane. 2018. "Proficiency Testing of Fingerprint Examiners with Bayesian Item Response Theory." *Law, Probability and Risk* 17 (2): 111–21.

Rasch, Georg. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

Fischer, Gerhard H, and Ivo W Molenaar. 2012. *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer Science & Business Media.

Lord, Frederic M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

van der Linden, Wim J, and Ronald K Hambleton. 2013. *Handbook of Modern Item Response Theory*. Springer Science & Business Media.

Boeck, Paul de, and Mark Wilson. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer.

Masters, Geoff N. 1982. “A Rasch Model for Partial Credit Scoring.” *Psychometrika* 47 (2): 149–74.

Luby, Amanda. 2019. *BlackboxstudyR: Fits Basic Irt Models to Fbi Black Box Data*.

Guo, Jiqiang, Jonah Gabry, and Ben Goodrich. 2018. *Rstan: R Interface to Stan*. https://CRAN.R-project.org/package=rstan.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. Third. Taylor & Francis.