Title: Accounting for Individual Differences among Decision-Makers with Applications in Forensic Evidence Evaluation

Presented by: Amanda S. Luby

Accepted by: Department of Statistics and Data Science

Readers:

Prof. Brian Junker, Advisor

Prof. Anjali Mazumder

Prof. Rebecca Nugent

Prof. Joseph B. Kadane

Dr. Reinoud Stoel

Approved by the Committee on Graduate Degrees:

Richard Scheines, Dean  Date
Carnegie Mellon University

Accounting for Individual Differences among Decision-Makers with Applications in Forensic Evidence Evaluation

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

In

Statistics

by

Amanda S. Luby

Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

August 2019
To my family.
Acknowledgements

First, thank you to Brian Junker, who agreed to take me on as a student late in the game without either one of us knowing what my dissertation topic would be. You’ve made me a much stronger researcher and writer, and I will always be grateful for the marathon meetings and drafts returned to me with more red ink than black. Thank you to my second advisor, Anjali Mazumder, who also played a huge role in shaping this dissertation. Thank you for introducing me to the broader world of forensic statistics and giving me opportunities to travel and share my work in places I never imagined were possible.

I’d also like to thank the rest of my committee: Rebecca Nugent, Jay Kadane, and Reinoud Stoel. Thank you to Rebecca for helping me keep the bigger picture in mind, being an invaluable mentor, and encouraging me to pursue every opportunity. I am so grateful for all of the advice over the past few years. Thank you to Jay for all of the conversations, feedback, editing, and for writing a conference paper with me that ultimately led to this project. Thank you to Reinoud for bringing your perspective to the project and helping me improve the arguments in the dissertation.

This dissertation wouldn’t have been possible without Steve Fienberg, who introduced me to statistical issues in forensic science and taught me the value of working on important problems that you care about.

I’d also like to thank all of the faculty and staff in the Statistics & Data Science department, particularly Robin Mejia and Bill Eddy and the other CSAFE affiliates who have provided feedback on my work. I am so grateful to have had the opportunity to learn from the best. Thank you to Margie, Jess, Kira, Christopher Peter, Sam, and Beth for helping with so much along the way.
Thank you to all of the wonderful PhD students in the department for your support and friendship. Thanks especially to my cohort-mates, office-mates, and roommates over the years including Maria Cuellar, Jackie Mauro, Brendan McVeigh, Purvasha Chakravarti, and Mikaela Meyer. Also thank you to everyone who has been a part of the Random Walkers legacy for the many hours of softball (and all of the other sports, I guess). Special shoutout to Shannon Gallagher and Kayla Frisoli for being the best co-organizers, co-presidents, co-authors, co-honorable-mention-winners, and friends.

Mom, Dad, Grandma, Taryn and Cait: I couldn’t have done it without you. Thank you so much for loving and supporting me unconditionally, even when I am grouchy or absent. Thank you to my parents, Doug and Jennifer, and my Grandma Sue Sue, for making education a priority for as long as I can remember and teaching me that ‘anything is possible’. Thank you all for the endless laughs and encouragement. I love you more than anything.

Lastly, thank you to Bret. You have been the most amazing supporter, co(dog)parent, chef, chauffeur, editor, and partner throughout this process. The past four years would have been infinitely more miserable without you. Time for the next adventure!
Abstract

Forensic science often involves the comparison of crime-scene evidence to a known-source sample to determine if the evidence arose from the same source as the reference sample. Common examples include determining if a fingerprint or DNA was left by a suspect, or if a bullet was fired from a specific gun. Even as forensic measurement and analysis tools become increasingly accurate and objective, final source decisions are often left to individual examiners’ interpretation of the evidence (President’s Council of Advisors on Science and Technology, 2016). The current approach to characterizing uncertainty in forensic decision-making has largely centered around the calculation of error rates, which is problematic when different examiners respond to different sets of items, as their error rates are not directly comparable. Furthermore, forensic analyses often consist of a series of steps. While some steps may be straightforward and relatively objective, substantial variation may exist in more subjective decisions.

The goal of this dissertation is to adapt and implement statistical models for human decision-making for the forensic science domain. Item Response Theory (IRT), a class of statistical methods used prominently in psychometrics and educational testing, is one approach that accounts for differences among decision-makers and additionally accounts for varying difficulty among decision-making tasks. By casting forensic decision-making tasks in the IRT framework, well-developed statistical methods, theory, and tools become available.

However, substantial differences exist between forensic decision-making tasks and standard IRT applications such as educational testing. I focus on three developments in IRT for forensic settings:
(1) modeling sequential responses explicitly, (2) determining expected answers from responses when an answer key does not exist, and (3) incorporating self-reported assessments of performance into the model. While this dissertation focuses on fingerprint analysis, specifically the FBI Black Box study [Ulery et al., 2011], methods are broadly applicable to other forensic domains in which subjective decision-making plays a role, such as bullet comparisons, DNA mixture interpretation, and handwriting analysis.
## Contents

1 Introduction

1.1 Preview of Dissertation

1.1.1 Background and Motivation: Application of Standard IRT to Forensic Decision-Making

1.1.2 Modeling Sequential Responses Explicitly

1.1.3 Determining Expected Answers from Responses Alone

1.1.4 Incorporating self-reported collateral responses

1.1.5 Summary, Future Work, and Broad Implications

2 Background and Motivation

2.1 Available Forensic Data

2.1.1 Proficiency Tests

2.1.2 Error Rate Studies

2.2 Item Response Theory

2.2.1 Binary Responses

2.2.2 Polytomous Responses

2.3 Application of standard IRT to forensic decision-making data

2.3.1 CTS Data

2.3.2 FBI Black Box Study
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4 Challenges Addressed in this Dissertation</td>
<td>23</td>
</tr>
<tr>
<td>3 Modeling Sequential Responses Explicitly</td>
<td>27</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>28</td>
</tr>
<tr>
<td>3.2 Item Response Trees</td>
<td>29</td>
</tr>
<tr>
<td>3.3 Decisions made in the FBI “Black Box” Study</td>
<td>31</td>
</tr>
<tr>
<td>3.4 Model Formulation</td>
<td>34</td>
</tr>
<tr>
<td>3.4.1 Ordered response tree</td>
<td>34</td>
</tr>
<tr>
<td>3.4.2 Binary decision process tree</td>
<td>37</td>
</tr>
<tr>
<td>3.4.3 Simplified scored tree</td>
<td>40</td>
</tr>
<tr>
<td>3.4.4 Prior Distributions</td>
<td>41</td>
</tr>
<tr>
<td>3.5 Results</td>
<td>42</td>
</tr>
<tr>
<td>3.5.1 Model Fit Comparison</td>
<td>42</td>
</tr>
<tr>
<td>3.5.2 Parameter estimates</td>
<td>44</td>
</tr>
<tr>
<td>3.5.3 Question Ambiguity vs Examiner Indecision</td>
<td>48</td>
</tr>
<tr>
<td>3.6 Discussion</td>
<td>53</td>
</tr>
<tr>
<td>3.7 Conclusions</td>
<td>55</td>
</tr>
<tr>
<td>4 Determining Expected Answers from Responses Alone</td>
<td>57</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>58</td>
</tr>
<tr>
<td>4.2 Literature</td>
<td>60</td>
</tr>
<tr>
<td>4.3 Model Formulation</td>
<td>61</td>
</tr>
<tr>
<td>4.3.1 Polytomous Ordered Response CCT (Latent Truth Rater Model)</td>
<td>61</td>
</tr>
<tr>
<td>4.3.2 Adapted LTRM as a Cumulative Logits Model</td>
<td>64</td>
</tr>
<tr>
<td>4.3.3 Adapted LTRM as an Adjacent Category Logits Model</td>
<td>65</td>
</tr>
<tr>
<td>4.3.4 IRTree for answer key generation</td>
<td>65</td>
</tr>
<tr>
<td>4.4 Sufficient Statistics</td>
<td>69</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>5.5.1 Joint Model for Standard IRT</td>
<td>103</td>
</tr>
<tr>
<td>5.5.2 Joint Model for IRTree</td>
<td>109</td>
</tr>
<tr>
<td>5.6 Discussion</td>
<td>113</td>
</tr>
<tr>
<td>5.7 Conclusions</td>
<td>115</td>
</tr>
<tr>
<td>5.8 Appendix</td>
<td>117</td>
</tr>
<tr>
<td>6 Summary, Future Directions, and Broader Implications</td>
<td>119</td>
</tr>
<tr>
<td>6.1 Dissertation Summary</td>
<td>119</td>
</tr>
<tr>
<td>6.2 Contributions</td>
<td>120</td>
</tr>
<tr>
<td>6.3 Future Work</td>
<td>121</td>
</tr>
<tr>
<td>6.3.1 Additional analyses on related data</td>
<td>121</td>
</tr>
<tr>
<td>6.3.2 Images as Covariates</td>
<td>125</td>
</tr>
<tr>
<td>6.3.3 Applications in Eyewitness Identification</td>
<td>126</td>
</tr>
<tr>
<td>6.3.4 Applications in Statistics Education Research</td>
<td>128</td>
</tr>
<tr>
<td>6.4 Conclusion</td>
<td>128</td>
</tr>
<tr>
<td>6.4.1 Results from the Black Box Study</td>
<td>129</td>
</tr>
<tr>
<td>6.4.2 Implications for Latent Print Examination</td>
<td>130</td>
</tr>
<tr>
<td>6.4.3 Broader Recommendations for Forensic Science</td>
<td>132</td>
</tr>
</tbody>
</table>
Chapter 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 our current criminal justice system, either a) there are no differences in the decision-making process between examiners, b) day-to-day forensic decision-making tasks are equally difficult, or c) human decision-making can be immediately removed from the process entirely.

The role of human decisions in forensic science is perhaps most studied in the fingerprint domain, which is the focus of this dissertation. High-profile examples of misidentification have inspired studies showing that fingerprint examiners, like all humans, may be susceptible to biased instructions [Dror and Rosenthal 2008] or influenced by external factors or contextual information [Dror et al. 2006; Dror and Cole 2010] and therefore unreliable in their final decisions. These studies contradict common perceptions of the accuracy of fingerprint examination [Thompson and Cole 2007; Garrett and Mitchell 2013; Thompson and Newman 2015] and demonstrate that fingerprint analysis is far from error-free.

After unknown fingerprint(s) from a crime scene have been collected (latent prints), a fingerprint from a known source (reference print) is found for comparison. The reference print may be taken from a known suspect or found through a database search. Once a suitable reference print has been found,
the examiner follows a process known as ACE-V (Analysis, Comparison, Evaluation, Verification) to determine whether the latent print and the reference print came from the same source. Each step in the ACE-V process is subjective, and examiners may differ in the value assessments they make, features on which they select to focus, and final source decisions. A better understanding of individual differences among examiners, and how those differences may impact performance, could lead to improvements in training programs and more accurate decisions (Luby and Kadane, 2018).

Although fingerprint examination is the focus of this dissertation, it is not the only forensic domain that relies on human decision-making. Firearm examination (See, e.g., National Research Council (2009) 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 (2009) pg. 163-167 on “Questioned Document Examination” and Mohammed (2019) for discussion) consists of examiners determining whether two samples of handwriting were authored by the same person, taking potential forgery or disguise into account. Interpreting mixtures of DNA evidence (PCAST 2016 Section 5.2) is a third example. 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 (Bright et al., 2019). The President’s Council of Advisors on Science and Technology (PCAST 2016) has recommended increased “black box” error rate studies for subjective forensic methods. While this dissertation focuses on statistical methods for analyzing the decision-making behavior of latent print examiners, the methods developed in further chapters are broadly applicable to the forensic science domain.

The FBI Black Box study (Ulery et al., 2011) was the first large-scale study performed to assess the accuracy and reliability of latent print examiners’ decisions. The items (fingerprint images) were designed to include a range of attributes and quality seen in casework and be 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, 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, only making 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 evaluated a different set of fingerprint images (Ulery et al. 2011, SI 3). In other words, it would be unfair to compare the error rate of someone who was given a set of “easy” items to the error rate of someone who was given a set of “difficult” items. 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, should likely be treated as mistakes.

Methods for analyzing forensic decision-making data should thus provide estimates for both participant proficiency and evidence difficulty, and these estimates should account for participants evaluating different sets of evidence. Item Response Theory (IRT), 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. There are, however, properties of forensic science applications which complicate an out-of-the-box application of IRT and require additional methodological developments. This dissertation casts forensic decision-making as an item-response task, identifies challenges in the straightforward application of IRT, and proposes solutions to each challenge.

1.1 Preview of Dissertation

The chapters of this dissertation are intended to be independent works, but they all share the theme of identifying challenges and proposing solutions in the application of IRT to forensic decision-making tasks.

1.1.1 Background and Motivation: Application of Standard IRT to Forensic Decision-Making

Chapter 2 provides an overview of the types of decision-making data that are collected in forensic science, an introduction to simple Item Response Models, and a demonstration of such Item Response Models applied to forensic settings. It thus defines forensic decision-making as a data-analysis task suitable for IRT analysis and lays out challenges in straightforward application. We build upon the challenges identified in this chapter in the remainder of the dissertation.

1.1.2 Modeling Sequential Responses Explicitly

Forensic decision-making tasks often include multiple assessments and decisions. For example, the current standard for latent print examination is known as ACE-V: Analysis, Comparison, Evaluation, and Verification. Each step in the ACE-V process consists of subjective decisions, and each result may be dependent on the examiner performing the analysis.

Chapter 3 aims to answer the following questions:

1. How can the loss of information through transforming multiple responses into one be avoided?
2. How does modeling differences in decision-making at each step in the process compare to traditional measures of proficiency?

3. Can the sequential nature of fingerprint analysis decisions be preserved?

### 1.1.3 Determining Expected Answers from Responses Alone

In forensic decision-making studies and proficiency tests, administrators generally know whether the evidence came from the same-source or different sources. However, analysts are often able to report an inconclusive instead of making a same-source or different-sources determination. In some cases, this may be the ‘correct’ answer due to a poor-quality piece of evidence or insufficient information. In other cases, the inconclusive may be ‘incorrect’, such as when there is sufficient information in the evidence to exclude the suspect but the examiner reports an inconclusive.

In casework, as opposed to research studies or proficiency tests, the source of the latent print is unknown. Generating known-source evidence evaluation tasks for assessment purposes is difficult and time-consuming. Methods to determine expected answers from responses alone would also allow for the creation of additional assessments using data collected from naturalistic settings, such as casework.

Chapter 4 addresses the following:

1. How can response data for which there is an unknown correct answer be modeled within the IRT framework?

2. When participants are allowed to specify an uncertain response, how can it be determined if that uncertainty is justified or a mistake, when there is no answer key for uncertain responses?

### 1.1.4 Incorporating self-reported collateral responses

Reported difficulty or confidence is often recorded alongside responses and used as a proxy for accuracy. These collateral responses provide a subjective measure of task challenge for each item. In casework, further investigative decisions may be made based on these collateral responses. Participants likely use
different thresholds for reporting such information. Individual differences among participants in self-reporting should be further studied to understand (a) how well-calibrated these self-reports are and (b) the implication of making casework decisions based on self-reports.

Intuitively, items that are rated as easier should have a higher proportion of correct responses, while items that are rated as more difficult should have a lower proportion of correct responses. Self-reported difficulty is therefore likely related to the responses and could be incorporated into a model for both.

Chapter 5 aims to answer the following questions:

1. Can modeling additional self-reports (e.g. reported difficulty) lead to a better understanding of the implications of making casework decisions based on such reports?

2. Can reported difficulty be incorporated directly into a model for responses?

3. Can responses be used to model differential use of the subjective reporting scales?

### 1.1.5 Summary, Future Work, and Broad Implications

Chapter 6 provides a brief summary of the work completed in this dissertation, future research directions, and implications of this dissertation for fingerprint examiners and the broader forensic science community.
Chapter 2

Background and Motivation

2.1 Available Forensic Data

The vast majority of forensic decision-making occurs in casework, which is not often made available to researchers due to privacy concerns or active investigation policies. Besides real-world casework, data on forensic decision-making is collected through proficiency testing and error rate studies. Proficiency tests are periodic competency exams that must be completed for forensic laboratories to maintain their accreditation, while error rate studies are research studies designed to measure casework error rates. As their names suggest, these two data collection scenarios serve completely different purposes. Proficiency tests are (currently) 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) responding to a small set of items (often < 20). Since every participant responds to every item, we can assess participant proficiency and item difficulty largely using the observed scores. As proficiency exams are designed to assess basic competency, most items are relatively easy and the vast majority of participants score 100% on each test. Error rate studies, on the other hand, rely on voluntary participation and therefore consist of a smaller
number of participants (fewer than 200), but use a larger pool of items (often 100 or more). The items are
designed to be difficult, and every participant does not respond to every item, which makes determining
participant proficiency and item difficulty a more complicated task.

2.1.1 Proficiency Tests

Proficiency testing companies include Collaborative Testing Services (CTS), Ron Smith and Associates
(RSA), Forensic Testing Services (FTS), and Forensic Assurance (FA). Each of these companies provides
different offerings for latent print examination. A summary of these differences is provided in Table 2.1
below.

<table>
<thead>
<tr>
<th></th>
<th>Tests/year</th>
<th>Q/test</th>
<th>Reports Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTS</td>
<td>2</td>
<td>10-12</td>
<td>Yes</td>
</tr>
<tr>
<td>RSA</td>
<td>2</td>
<td>10</td>
<td>Yes</td>
</tr>
<tr>
<td>FTS (No latent print offerings)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FA</td>
<td>2</td>
<td>unknown</td>
<td>No</td>
</tr>
</tbody>
</table>

CTS Data

CTS has provided copies of twelve proficiency exams distributed semi-annually between 2012 and 2017.
Each proficiency exam consists of eleven or twelve items, where each item corresponds to an unknown
latent print (e.g. Figure 2.1b). Participants are tasked with determining the source of the print from
a pool of four known donors (e.g. Figure 2.1a), if any. Each of the twelve tests was taken by 300-500
participants.

Results from these tests are available on the CTS website, with the following warning:

Since these participants are located in many countries around the world, and it is their option how
the samples are to be used (e.g. training exercise, known or blind proficiency testing, research and
development of new techniques, etc.) the results compiled in the Summary Report are not intended to
be an overview of the quality of work performed in the profession and cannot be interpreted as such.
Additionally, since the test environment is not controlled, it is impossible to determine whether responses correspond to an individual examiner’s decision or whether responses represent the consensus answer of a group of examiners working together on the exam.

![Figure 2.1: Examples of latent and reference samples provided in CTS proficiency exams.](image)

Even with the above limitations, the CTS data provides valuable insight into examiner performance. Since laboratories must participate in proficiency tests on a regular basis in order to maintain accreditation, the results from these tests likely provide a more representative sample of practicing examiners than optional research studies. Proficiency test results also provide longitudinal data across different years, and access to the images alongside the item responses allow for additional inferences to be made.

### 2.1.2 Error Rate Studies

AAAS (2017) identified twelve existing error rate studies in the latent print domain, and a summary of those studies is provided here. The number of participants (N), number of items (I), false positive rate, false negative rate, and reporting strategy varies wildly across the studies and are summarized in Table 2.2 below. There are, however, problems or discrepancies noted in many of the studies. Evett and Williams (1996) did not report inconclusives, making results difficult to evaluate relative to the
other studies. Wertheim et al. (2006) reported that many of the errors were caused by “clerical error” rather than a false decision, suggesting design flaws. Gutowski (2006) presented data from the CTS exams, rather than designing new items. Langenburg et al. (2009) split participants into three groups and attempted to introduce contextual bias in two of the groups, meaning the reported error rates were only for the control group, or 15 of 43 examiners. Langenburg (2009) report results from a pilot study in which there were very few conclusive results from different-source prints, which could be a contributing factor to the lack of false positives. The false negative rate is reported for two procedures: one that included verification and one that did not. Tangen et al. (2011) asked participants to rate the likelihood that two fingerprints came from the same source on a 12-point scale, then labeled scores 1-6 as identifications and 7-12 as exclusions, which is not the typical reporting practice for latent print analysis. Ulery et al. (2011) is generally regarded as the most well-designed error rate study for latent print examiners (AAAS 2017, PCAST 2016). Ulery et al. (2012) tested the same examiners on 25 of the same items they were shown seven months earlier, and found that the repeatability of all comparisons decisions combined was 90% for same-source pairs, and 85.9% for different-source pairs. Kellman et al. (2014) required examiners to make a determination about the source of a latent print in only three minutes, likely leading to larger error rates. Pacheco et al. (2014) reported that incorrect decisions may have arisen due to clerical error. Liu et al. (2015) selected prints to represent “difficult cases”, and conclude that “close nonmatch” pairs found by an AFIS (Automated Fingerprint Identification System) search may lead to increasing false positive rates in casework.

<table>
<thead>
<tr>
<th>Study</th>
<th>N</th>
<th>I</th>
<th>False Pos</th>
<th>False Neg</th>
<th>Inconclusive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evett and Williams (1996)</td>
<td>130</td>
<td>10</td>
<td>0</td>
<td>0.007%</td>
<td>Not reported</td>
</tr>
<tr>
<td>Wertheim et al. (2006)</td>
<td>108</td>
<td>10</td>
<td>1.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Langenburg et al. (2009)</td>
<td>15 (43)</td>
<td>6</td>
<td>2.3%</td>
<td>7%</td>
<td></td>
</tr>
<tr>
<td>Langenburg (2009)</td>
<td>6</td>
<td>120</td>
<td>0</td>
<td>0.7%/2.2%</td>
<td></td>
</tr>
<tr>
<td>Langen et al. (2011)</td>
<td>37 (74)</td>
<td>36</td>
<td>0.0037</td>
<td>Not allowed</td>
<td></td>
</tr>
<tr>
<td>Ulery et al. (2011)</td>
<td>169</td>
<td>744 (100)</td>
<td>0.17%</td>
<td>7.5%</td>
<td></td>
</tr>
<tr>
<td>Ulery et al. (2012)</td>
<td>72</td>
<td>744 (25)</td>
<td>0</td>
<td>30% of previous</td>
<td></td>
</tr>
<tr>
<td>Langenburg et al. (2012)</td>
<td>159</td>
<td>12</td>
<td>2.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kellman et al. (2014)</td>
<td>56</td>
<td>200 (40)</td>
<td>3%</td>
<td>14%</td>
<td>Not allowed</td>
</tr>
<tr>
<td>Pacheco et al. (2014)</td>
<td>109</td>
<td>40</td>
<td>4.2%</td>
<td>8.7%</td>
<td></td>
</tr>
<tr>
<td>Liu et al. (2015)</td>
<td>40</td>
<td>5</td>
<td>0.11%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: Summary of existing studies that estimate error rates in latent print examination
The FBI “black box” dataset is freely available from the FBI. Each row in the data file corresponds to an examiner × item response. In addition to the Examiner ID and item Pair ID, the data contains:

- **Mating**: whether the pair of prints were “Mates” (a match) or “Non-mates” (a non-match)

- **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 to 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”

[https://www.fbi.gov/services/laboratory/scientific-analysis/counterterrorism-forensic-science-research/black-box-study-results](https://www.fbi.gov/services/laboratory/scientific-analysis/counterterrorism-forensic-science-research/black-box-study-results)


2.2 Item Response Theory

For $P$ individuals responding to $I$ test items, one can express the binary responses (i.e. correct/incorrect) as a $P \times I$ matrix, $Y$. Item Response Theory (IRT) is based on the idea that the probability of a correct response depends on individual proficiency, $\theta_p$, $p = 1, \ldots, P$, and item difficulty, $b_i$, $i = 1, \ldots, I$.

2.2.1 Binary Responses

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 in this dissertation. The probability of a correct response is modeled as the logistic function of the difference between the participant proficiency, $\theta_p$ ($p = 1, \ldots, P$), and the item difficulty, $b_i$ ($i = 1, \ldots, I$):

$$P(Y_{pi} = 1) = \frac{1}{1 + \exp(-(\theta_p - b_i))}.$$  (2.1)

To identify the model, the convention of constraining the mean of the participant parameters ($\mu_\theta$) to be equal to zero is used. 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 $b_i < 0$ item $i$ is an “easier” item and the average participant is more likely to answer item $i$ correctly. If $b_i > 0$ then item $i$ is a more “difficult” item and the average participant is more likely to answer item $i$ incorrectly. Other common conventions for identifying the model include setting a particular $b_i$ or the mean of the $b_i$s equal to zero.

The item characteristic curve (ICC) describes the relationship between proficiency and performance on a particular item (see Figure 2.2 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 2.1 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 $\theta_p$ and $b_i$, the identifiability problem is solved. Using the convention above, we modeled $\theta_p \sim N(0, \sigma^2_\theta)$ and $b_i \sim N(\mu_b, \sigma^2_b)$, although more complicated prior distributions are certainly possible.

**Extensions**

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 de Boeck and Wilson (2004).

**2.2.2 Polytomous Responses**

Item response 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.
The partial credit model (PCM) [Masters 1982] is especially useful for modeling partially correct responses, although may be applied in other contexts where the responses can be ordered. When $Y_{pi}$ is binary, the partial credit model is equivalent to the Rasch model. Under the PCM, the probability of response $Y_{pi}$ depends on $\theta_p$, the proficiency of participant $p$ as in the above models; $m_i$, the maximum score for item $i$ (and the number of step parameters); and $\beta_{il}$, the $l^{th}$ step parameter for item $i$ ($l = 0, \ldots, m_i$):

\[
P(Y_{pi} = 0) = \frac{1}{1 + \sum_{k=1}^{m_i} \exp \sum_{l=1}^{k}(\theta_p - \beta_{il})} \tag{2.2}
\]

\[
P(Y_{pi} = y, y > 0) = \frac{\exp \sum_{l=1}^{y}(\theta_p - \beta_{il})}{1 + \sum_{k=1}^{m_i} \exp \sum_{l=1}^{k}(\theta_p - \beta_{il})}. \tag{2.3}
\]

The PCM is therefore an adjacent-categories logistic regression model. An example PCM is shown in Figure 2.3 by plotting the probabilities of observing each of three categories as a function of $\theta_p$ (analagous to the ICC curves above).
Rating Scale Model

The Rating Scale Model (RSM, Andrich (1978)) is also an adjacent-categories polytomous response IRT model, like the PCM, but places an additional restriction on the item parameters. It is assumed that the difference between item step parameters, e.g. $\beta_i - \beta_{i-1}$ in Equation 2.3, are constant across all items. That is,

$$\beta_{i,1} = \beta_i,$$

$$\beta_{i,k} = \beta_{k-1} + \lambda$$

where $\lambda$ is the difference between categories 1 and 2 on the latent scale.

Graded Response Model

The Graded Response Model (GRM, Samejima (1969)) is a cumulative-categories version of a polytomous response model. It can also be interpreted as a proportional odds model.

$$P(Y_{pi} \geq k) = \frac{\exp(\theta_p - \beta_{ik})}{1 + \exp(\theta_p - \beta_{ik})}$$

Note that a rating scale version of the GRM can also be constructed by restricting $\beta_{ik}$ as in Equation 2.4 above.
2.3 Application of standard IRT to forensic decision-making data

2.3.1 CTS Data

Luby and Kadane [2018] provide the first IRT analysis of forensic data, using responses from a single CTS exam (Latent Print Examination Test Number 16-515/516). The majority of participants ($n = 383$ of 431) correctly answered all twelve of the items. Eleven of the twelve items had over 98% correct response rates, with one item (Q2) having a 100% correct response rate.

Using IRT, we showed that estimated proficiency and difficulty are consistent with the observed scores for each person and item. We identified one ambiguous item (Q6) in which the unknown latent print came from an index finger joint that is visible only on the palm print card (and not on the ten-print card). The consensus answer was thus ‘Person D, right index finger’, but ‘Person D, right palm’ could also be scored as correct. Many participants who reported ‘Person D, right index finger’ noted that they used the right palm print of Person D to make the identification. We showed that results were sensitive to how this response was scored (Figures 2.4, 2.5, and 2.6).

We also provided a brief simulation showing hypothetical results of a mixed group of novices and experts on (a) the CTS exam, and (b) a hypothetical exam with more difficult items. Results demonstrated that an exam with more difficult items would be better at distinguishing between the two groups.

![Figure 2.4: Item Characteristic Curve for one CTS exam, treating ‘Person D, right index finger’ as the correct answer for Q6.](http://www.ctsforensics.com/assets/news/3616_Web.pdf)

![Figure 2.5: Item Characteristic Curve for one CTS exam, treating ‘Person D, right palm’ as the correct answer for Q6.](http://www.ctsforensics.com/assets/news/3616_Web.pdf)

![Figure 2.6: ICC for one CTS exam, treating both ‘Person D, right index finger’ and ‘Person D, right palm’ as correct answers for Q6.](http://www.ctsforensics.com/assets/news/3616_Web.pdf)
2.3.2 FBI Black Box Study

Unlike the CTS example above, the Black Box Study does not provide responses scored as correct/incorrect. An example IRT analysis of the Black Box Study is provided in Luby (2019), which we summarize here.

In order to fit an IRT model to the Black Box Study, the responses must be scored. Responses should be 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 should be scored as incorrect if they are false identifications (Mating == Non-mates and Compare_Value == Individualization) or exclusions (Mating == Mates and Compare_Value == Exclusion).

It is unclear, however, how the inconclusive responses should be scored. “Inconclusive” is never keyed as the correct response, so there is no clear way to tell when an inconclusive should have been expected. There are a large number of inconclusive answers (4907 of 17121 responses), and examiners vary on which latent print pairs are inconclusive.

Possible scoring schemes include:

• ‘Inconclusive MCAR’: all inconclusive responses are treated as missing

• ‘Inconclusive Incorrect’: all inconclusive responses are treated as incorrect

• ‘Partial Credit’: inconclusive responses are treated as partial credit

• ‘Consensus-based’: if the reason provided for the inconclusive matches the consensus reason, it is scored as correct, otherwise the inconclusive is scored as incorrect

In Ulery et al. (2011), the inconclusives were treated as missing completely at random (‘Inconclusive MCAR’) when computing error rates. The Rasch model was fit using a Bayesian framework under the ‘inconclusive MCAR’ scheme using Stan. Figure 2.7 shows estimated proficiencies of examiners when responses are scored as such, with 95% posterior intervals, 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. One of the participants who made a false positive error still received a relatively high proficiency estimate due to having a small false negative rate.

![Diagram showing estimated proficiency by observed false positive rate (left) and false negative rate (right). Participants who made at least one false positive error, i.e. the nonzero cases in the left-hand plot, are colored in blue on the right-hand plot.](image)

**Figure 2.7:** Estimated proficiency by observed false positive rate (left) and false negative rate (right). Participants who made at least one false positive error, i.e. the nonzero cases in the left-hand plot, are colored in blue on the right-hand plot.

The estimated proficiencies correlate with the observed score (Figure 2.8). That is, participants with a higher observed score generally received larger proficiency estimates than participants with lower scores. There were, however, cases where participants scored roughly the same on the study but received vastly different proficiency estimates. For example, the highlighted participants in the right-side plot above all scored between 94% and 96%, but their estimated proficiencies ranged from $-1.25$ to 2.5.

If those participants who scored between 94% and 96% are examined more closely, the discrepancies in their proficiencies can largely be explained by the difficulty of the specific set of items they responded to. This is evidenced by the positive trend in the right-side plot of Figure 2.8. In addition to the observed score and difficulty of the item set, the other factor that plays a role in the proficiency estimate is the number of items the participant answers conclusively (i.e. individualization or exclusion). When the inconclusive responses are treated as missing, participants who are conclusive more often generally receive higher estimates of proficiency than participants who are conclusive less often.

Proficiency estimates, however, are dependent on the scoring method chosen. In the original black box study, inconclusive responses were treated as missing when calculating error rates because those
decisions would not lead to further investigation or complete exclusion of a suspect in casework. One could instead make the argument that inconclusive responses should be scored as incorrect if other examiners were able to make a conclusive decision, or that inconclusive responses should be treated as a distinct third category rather than correct, incorrect, or missing.

To illustrate the difference in results between different scoring methods, the dataset was scored in two additional ways: using the consensus inconclusive reason for each item as the correct answer for inconclusives (‘consensus-based’) and treating inconclusive decisions as ‘in-between’ an error and a correct response (‘partial credit’) [Luby, 2019]. The proficiency estimates are compared to the observed scores for each participant under each of the three scoring schemes in Figure 2.9. 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 \((\text{#Correct} + \text{#Inconclusive})/(2 \times \text{#Responses})\) to scale the score to be between 0 and 1.

Treating the inconclusives as missing completely at random (‘inconclusive MCAR’) leads to both the smallest range of observed scores and largest range of estimated proficiencies. Harsher scoring methods (e.g. ‘consensus-based’ and ‘partial credit’) do not necessarily lead to lower estimated proficiencies. For instance, the participants who scored around 45% under the consensus-based scoring method received
higher proficiency estimates than the participant who scored 70% under the inconclusive MCAR scoring method. Harsher scoring methods lead to more items being estimated as difficult, and participants who get harder items correct are estimated to be more proficient than those who get easier items correct. Thus lower observed scores under harsher scoring schemes often lead to higher proficiency estimates than higher observed scores under more lenient scoring schemes.

Also note that the uncertainty intervals under the inconclusive MCAR scoring scheme were noticeably larger than under the other scoring schemes. All of the inconclusives, nearly a third of the data, were treated as missing, which 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 item) and therefore less uncertainty in the proficiency estimates.

The range of proficiency estimates 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, consider setting the mastery threshold at $-0.5$. Examiners then have not demonstrated mastery if the upper limit of their proficiency uncertainty estimate is below $-0.5$, illustrated in the right plot of Figure 2.9.

The number of examiners that did not demonstrate mastery varies based on the scoring method used (11 for consensus-based, 8 for partial credit, and 11 for inconclusive MCAR) due to the variation in ranges of proficiency estimates. Additionally, for each of the scoring schemes, there were 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 items are given higher proficiency estimates than participants that answered the same number of easier items.

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

The plot on the left in Figure 2.10 shows both a change in examiner proficiencies across scoring schemes (i.e. the lines connecting the proficiencies are not horizontal) as well as a change in the ordering of examiner proficiencies (i.e. the lines cross one another). In other words, different scoring schemes affect examiner proficiencies in different ways.

The plot on the right in Figure 2.10 highlights five participants who saw substantial changes in their proficiency estimates under different scoring schemes. Examiners 105 and 3 benefited from the leniency in scoring when inconclusives were treated as missing (inconclusive MCAR). When inconclusives were possibly scored as incorrect (consensus-based) or partial credit (partial credit), they saw a substantial decrease in their proficiency due to reporting a high number of inconclusives and differing from other examiners in their reasoning for those inconclusives. Examiners 142, 60 and 110, on the other hand, were hurt by the leniency in scoring when inconclusives are treated as missing (inconclusive MCAR).
Their proficiency estimates increased when inconclusives were scored as correct when they matched the consensus reason (consensus-based) or were worth partial credit.

![Estimated Proficiency](chart)

**FIGURE 2.10:** Change in proficiency for each examiner under the three scoring schemes. The right side plot has highlighted five examiners whose proficiency estimates change the greatest among the three schemes.

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 items seen. Additionally, IRT implicitly accounts for the inconclusive rate of different participants and provides estimates of uncertainty for both participant proficiency and item difficulty.

Three scoring schemes were presented above, 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.

Even if the ‘true’ scoring scheme for inconclusives were available, there are also ‘no value’ decisions to contend with. The first decision participants were asked to make was whether the latent print had enough value to move forward. If they decided the latent print had no value, they never saw
the reference print. Similar to inconclusive responses, the ‘no value’ decisions were prevalent and varied across examiners, and characterizations of examiner proficiency should take these tendencies into account.

2.4 Challenges Addressed in this Dissertation

The remainder of this thesis addresses the methodological challenges associated with IRT analyses of complex decision-making data, with a focus on analyzing the Black Box Study. We have identified three challenges in applying IRT to the Black Box study which are all illustrated in Figure 2.11. Each panel represents an item that appeared in the study, and the ribbons represent participants who responded to that item. Each panel begins with the latent evaluation (Has value or No value (NV)), followed by the source decision (Exclusion (Exc), Individualization (ID), Inconclusive (Inc.) or no value (NV)), and finally the reported difficulty (A represents the easiest, while E represents the most difficult).

![Figure 2.11: Responses to a low-consensus item (left) and high-consensus item (right), illustrating (1) disagreements in both latent evaluation and source decisions, (2) the difficulty in determining when an inconclusive or no value response is expected, and (3) substantial variation in reported difficulty, even when there is agreement in preceding decisions.](image)

Item ‘M003064’ (shown on the left) is a true match, but there is substantial disagreement among examiners regarding both latent value (furthest left category) and source decision (middle category). Examiners who responded to item ‘N057413’, on the right, largely agreed about both the latent value and source decision. This suggests that there are both examiner and item effects at each step in the process.
Treating responses as correct or incorrect (or partial credit) obscures some of these effects, and we may instead prefer an approach that maintains the structure of the responses. A method for preserving the structure of these sequential responses is introduced in Chapter 3.

While it is known whether each item showed above is a true match or a true non-match, there is no designation provided by the FBI when an inconclusive or a no value is expected. Instead, examiner consensus can be used to determine expected responses for each item. In the example above, the item on the right is a high-consensus item, and from the graph alone it is clear that an ‘exclusion’ is likely the expected answer. The item on the left, on the other hand, is not so clear. It appears that ‘inconclusive’ was the most common response, but there are also a number of ‘no value’ and ‘individualization’ (which is technically correct) conclusions. In Chapter 4 IRT approaches for data without an answer key are compared.

Finally, Figure 2.11 illustrates substantial variability in reported difficulty, even when there is agreement in preceding decisions. This suggests that, much like the responses themselves, there are both examiner and item effects in reported difficulty. Reported difficulty in this case can be thought of as ‘collateral information’: they are not the purpose of the assessment, but are related to the primary response of interest and collected at the same time. In Chapter 5 a joint modeling approach is introduced for responses and collateral information which use the same latent variables. This approach is useful not only for reported difficulty in the black box study, but could also be used for additional types of collateral information that arise in forensic science, such as time spent on the analysis, confidence, or a subjective likelihood ratio.

Chapters 3, 4, and 5 therefore each address a specific challenge that arises in the Black Box study: modeling sequential responses explicitly, determining expected answers from responses alone, and incorporating self-reported collateral information, respectively. Although the focus of the dissertation is on latent print examination, and the Black Box study in particular, the methods proposed would be widely applicable to other complex decision-making scenarios if appropriate data were collected.
Chapter 6 provides a discussion of implications for the Black Box study, latent print examination, and forensic science; as well as lays out future directions for this line of research.
Chapter 3

Modeling Sequential Responses

Explicitly

Most Item Response Models were designed for binary scored responses (correct/incorrect) or other standard response schemes consisting of a single measurement (e.g. Likert scale responses). In forensic decision-making tasks, however, it is not clear how to collapse the decision-making process, which often includes multiple assessments and decisions, into a single binary response. Any such ‘scoring’ of the data will result in some loss of information. This chapter seeks to answer:

1. How can collapsing multiple responses into one (and therefore losing information) be avoided?

2. Can the sequential nature of decisions be preserved?

3. How does modeling differences in decision-making at each step in the process compare to traditional measures of proficiency?
3.1 Introduction

The current standard procedure for latent print examination is known as ACE-V: Analysis, Comparison, Evaluation, and Verification. The verification step involves a second examiner verifying the decisions of the first, and was not addressed in the Black Box study. The analysis step corresponds to assigning a value to the latent print, but the comparison and evaluation steps are indistinguishable from one another in the data. While a model that assesses individual differences at each step in the ACE-V process would be useful, such a model cannot be constructed using the available data. Instead of assessing variation in the ACE-V process directly, individual differences in each decision that was recorded in the Black Box study can be assessed.

Each response in the Black Box data can be separated into three distinct decisions, where each decision is dependent on the preceding decisions. The first decision that participants were tasked with is to assess the quality of the latent print. If the latent print is deemed too poor, participants will not be shown the reference print and will not proceed with further decisions. The second decision made is the source evaluation of the pair of fingerprints (latent and reference), which is recorded as either an individualization (i.e. same-source), exclusion (different sources), or inconclusive. Additionally, reasons were recorded for inconclusive and exclusion decisions. Since each response can be represented as a series of sequential decisions, any reformulation of the recorded data to fit within a traditional IRT model necessarily results in the loss of information.

In traditional IRT applications (such as educational testing), the correct answer for any given item is often known. When comparing latent prints, on the other hand, analysts may report an inconclusive instead of a match or non-match. In some cases, this may be the ‘correct’ answer due to an insufficient amount of information in the latent print (e.g. very poor quality or a partial print). In other cases, the inconclusive may be an ‘incorrect’ answer, such as when there is sufficient information in the evidence to exclude the suspect as the source but the examiner reports an inconclusive (see, e.g. Ulery et al. (2014) for discussion of sufficiency in latent prints). As evidenced in Chapter 2, inconclusive responses complicate the scoring of the black box study due to both their prevalence and ambiguity. Examiners
vary in determining which latent print pairs are of no value and which are inconclusive. Since correct answers are not keyed by the FBI, these responses lead to an ambiguity in scoring. We would like to separate examiner effects from item effects in inconclusive responses, and determine how they interact to result in inconclusive responses.

The Item Response Trees (IRTrees, De Boeck and Partchev 2012) framework provides a solution for modeling each decision explicitly. IRTrees represent responses with decision trees where branch splits represent hypothesized internal decisions and leaves are observed outcomes. Sequential decisions can be represented explicitly in the IRTree framework, and node splits need not represent scored decisions. IRTrees, however, can be constructed in a variety of ways. We therefore provide a way to think about model fit in this setting, including traditional measures of model fit (such as predictive performance) as well as the usefulness/interpretability of parameters.

IRTrees are introduced more fully in Section 3.2. The application and data set are detailed in Section 3.3. Section 3.4 describes the formulation of three possible IRTree models for the application at hand, and Sections 3.5 and 3.6 present results and discussion, respectively.

### 3.2 Item Response Trees

Item Response Trees (IRTrees, De Boeck and Partchev 2012) use decision trees to describe hypothesized cognitive processes, where the leaves are the final observed outcome. IRTrees can be linear (at least one branch from each internal node leads to a terminal node), nested (all branches from an internal node lead to other internal nodes), or a combination of the two (Jeon and De Boeck 2016). They can also be binary (each internal node is a choice between two branches) or polytomous (an internal node is a choice between more than two branches). The IRTree formulation can thus represent a wide variety of response formats and response processes, easily adapted for binary responses, unipolar scales, bipolar scales, and Likert responses. They can also model unordered responses in some cases (Lopez-Sepulcre et al. 2015). IRTrees have been used in applications such as differentiating fast and slow intelligence.
response styles in multiple-choice items (Plieninger and Meiser, 2014), modeling answer change behavior (Jeon et al., 2017), and response behavior in ecological data (Lopez-Sepulcre et al., 2015).

Figure 3.1 illustrates a linear IRTree with three possible outcome categories (e.g. $Y = 1, 2, 3$). $Y_1^*$ and $Y_2^*$ are nodes constructed to represent internal decisions that lead to each of three outcomes. If $Y_{ij}$ denotes the response of participant $i$ ($i = 1, ..., P$) to item $j$ ($j = 1, ..., I$), we’ll use $Y_{1ij}^*$ to denote the choice of left or right branch at $Y_1^*$ for person $i$ at item $j$.

The probability of choosing the left branch at node 1 ($Y_1^*$ in Figure 3.1) can be modeled using standard Item Response Theory. The Rasch model was used for interpretability and computational convenience: $P(Y_{1ij}^* = 1) = \logit^{-1}(\theta_{1i} - b_{1j})$, where $\theta_{1i}$ denotes the latent trait involved with choosing the left branch for person $i$ and $b_{1j}$ is the corresponding Rasch parameter for item $i$. In a standard Rasch model for correct/incorrect outcomes, the item parameters ($b_i$) correspond to the difficulty of the item, where higher values of $b_i$ decrease the probability that $Y_{ij} = 1$. When IRTree branch decisions do not correspond to incorrect/correct choices, the item parameters represent an “item tendency” towards one branch over the other rather than difficulty.

The model for the probability of choosing the left branch at $Y_2^*$ is similar, except it is conditional on $Y_1^*$ being equal to zero (i.e. we model $P(Y_{2ij}^* = 1|Y_{1ij}^* = 0)$ instead of $P(Y_{2ij} = 1)$). The probability of each observed response (Outcome 1, 2, or 3) is then the product of the probabilities of the internal branches leading to each leaf in the tree. For the example given in Figure 3.1.
\[ P(Y_{ij} = \text{Outcome 1}) = P(Y^*_{ij} = 1) \]  
\[ P(Y_{ij} = \text{Outcome 2}) = P(Y^*_{1ij} = 0) \times P(Y^*_{2ij} = 1|Y^*_{1ij} = 0) \]  
\[ P(Y_{ij} = \text{Outcome 3}) = P(Y^*_{1ij} = 0) \times P(Y^*_{2ij} = 0|Y^*_{1ij} = 0) \]

Formulating an item response model for categorical outcomes with a tree structure is especially useful for modeling complex responses for the following reasons: 1) it does not impose an explicit ordering on the categories when the tree is binary, 2) it does not assume the same latent trait governs the entire decision-making process, 3) it does not require responses to be keyed as “correct” or “incorrect”.

### 3.3 Decisions made in the FBI “Black Box” Study

Although the purpose of the Black Box study was to estimate false positive and false negative error rates, the recorded data also contains additional information about examiners’ decision-making process. Recall that each recorded response to an item consists of three decisions:

1. Value assessment for the latent print only

2. Source assessment of the latent/reference pair

3. (If inconclusive) Reason for inconclusive

Note that for the value assessment, we do not distinguish between ‘value for individualization’ and ‘value for exclusion only’ for this analysis, and treat the value assessment as a binary response (‘Has value’ vs ‘No value’) instead. There were 40 conflicting responses, made by 32 unique participants, in which a participant reported that a latent print had ‘value for exclusion only’ and then proceeded to make an individualization for the second decision. As Haber and Haber (2014) note, only 17% of participants reported that they used ‘value for exclusion only’ in their normal casework on a post-experiment questionnaire, and participants may have interpreted this decision in different ways. These
discrepancies have been accounted for by treating the value evaluation as a binary response – either ‘has value’ or ‘no value’.

Latent print examiners have been found to vary in their tendencies towards ‘no-value’ and ‘inconclusive decisions’ (Ulery et al., 2011). Figure 3.2 shows the distribution of the number of inconclusive and no value decisions reported by each examiner. Although most examiners report 20-40 inconclusives and 15-35 ‘no value’ responses, some examiners report as much as 60 or as few as 5. By modeling these responses explicitly within the IRTree framework, not only can individual differences in proficiency among examiners be assessed, but also differences in tendency towards value assessments and inconclusive responses.

Furthermore, this work has taken a novel approach to modeling this data by constructing a bipolar scale from the possible responses. This not only provides an ordering for the responses within each sub-decision (i.e. source decision and reason for inconclusive), but allows the sub-decisions to be combined in a logical way. This scale is also consistent with other hypothetical models for examiner decision-making (Dror and Langenburg, 2019). Based on the description of each option for an inconclusive response, the ‘Close’ inconclusives are more similar to an individualization than the other inconclusive reasons. The ‘No overlap’ inconclusives are more similar to exclusions than the other inconclusive reasons, under the assumption that the reference prints are relatively complete. That is, if there are no overlapping areas between a latent print and a complete reference print, the two prints likely came from
different sources. The ‘insufficient’ inconclusives are treated as the center of the constructed match/no-match scale. This scale is represented in Figure [3.3]

![Figure 3.3: FBI black box responses as a bipolar scale.](image)

Inconclusive results are far more likely to be reported for true matches as opposed to true non-matches in this data (Table [3.1]). This may be due to a difference in the way examiners assess certainty for different types of stimuli, for instance, requiring a higher degree of certainty to report an individualization than to report an exclusion, or to an underlying difference in difficulty between true matches and non-matches in the question pool. The distribution of inconclusive reasons is also different between true matches and true non-matches.

| Table 3.1: Breakdown of inconclusive reason by whether the question was a true match or non-match. |
|-------------------------------------------------|-------------------------------------------------|
| True Match | Non-Match |
| Close | 804 | 212 |
| Insufficient | 894 | 94 |
| No Overlap | 2177 | 726 |

There are many adjustments that could be made to the scale proposed in Figure [3.3]. For instance, ‘No Overlap’ may belong in the middle of the scale, using the reasoning that if there is no overlapping area between the latent and reference print there is no information available to make a decision.

This scale also provides a connection between source-conclusion reporting (used in the Black Box study) and the likelihood-ratio (LR) based approach method of reporting, which is used throughout Europe and Canada. In the LR reporting paradigm, source conclusions are reported as:

\[
\frac{P(\text{Evidence}|\text{Same Source})}{P(\text{Evidence}|\text{Different Sources})}
\]  

\[ (3.4) \]
The ‘individualization’ side of the scale corresponds to $LR \to \infty$, while the ‘exclusion’ side of the scale corresponds to $LR \to 0$. This connection is discussed further in Chapter 6.

### 3.4 Model Formulation

In the IRTree model formulation, each internal node split in the tree (e.g., $Y^*_1$, $Y^*_2$, etc.) is modeled with an IRT model. The number of internal nodes in the tree then determines the number of latent traits to be estimated in the model. Participant $i$’s response to item $j$ is denoted as $Y^*_{ij}$ and internal node outcomes as $Y^*_{kij}$ (for $k$ internal nodes).

We will compare three IRTree models – one that directly represents the observed responses as ordered categories, one that represents a hypothesized binary internal decision process, and one that scores the responses to fit a more traditional Rasch model. Each of the three models presented below have different strengths and weaknesses, which are explored throughout the remainder of this section.

#### 3.4.1 Ordered response tree

The internal nodes in the ordered response tree directly correspond to each recorded decision in the black box study. The first internal node ($Y^*_1$) thus represents the value assessment, the second internal node ($Y^*_2$) represents the source assessment, and if an inconclusive response is observed the third internal node ($Y^*_3$) represents the inconclusive reason. The tree structure for the response-based tree is shown in Figure 3.4.

Note that for the internal nodes with polytomous responses ($Y^*_2$ and $Y^*_3$), it is assumed that the possible responses are ordered in order to model each decision using a single latent trait. For example, it is assumed that individualization $<$ inconclusive $<$ exclusion for $Y^*_2$, and so the latent participant parameter represents a tendency towards exclusion. This particular ordering of responses is can be seen in the bipolar scale constructed in Figure 3.3. It should be noted that participants were not provided
with this scale and may not have been thinking about an order to the responses, yet this ordering may be a useful way to organize responses.

\[ Y_1^* \] is parameterized using a Rasch model. Possible polytomous-response model formulations for \( Y_2^* \) and \( Y_3^* \) include the graded response model (Samejima [1969], rating scale model (Andrich [1978]), and partial credit model (Masters [1982]). The partial credit model was chosen here because it allows for the steps between categories (e.g. moving from Inconclusive to Exclusion for \( Y_2^* \)) to vary across items.

The probability of observing each of the possible responses can then be calculated by multiplying the branch probabilities that lead to each leaf. The probability of observing a ‘No value’ response is simply the Rasch model corresponding to the first split:

\[
P(Y_{ij} = \text{No Value}) = P(Y_{1ij} = 1)
\]

\[
= \logit^{-1}(\theta_{1i} - b_{j1}).
\]
The probability of observing an ‘Individualization’ or an ‘exclusion’ consists of both a Rasch probability (corresponding to $Y_{1i}^*$) and a partial credit model probability (corresponding to $Y_{2i}^*$):

$$P(Y_{ij} = \text{Individualization}) = P(Y_{1ij}^* = 0) \times P(Y_{2ij} = 1 | Y_{1ij} = 0)$$

$$= (1 - \logit^{-1}(\theta_{1i} - b_{j1})) \times \frac{\exp(\theta_{2i})}{\sum_{l=0}^{2} \exp \sum_{m=0}^{l} (\theta_{2i} - b_{jm})}$$

(3.7)

and

$$P(Y_{ij} = \text{Exclusion}) = P(Y_{1ij}^* = 0) \times P(Y_{2ij} = 3 | Y_{1ij} = 0)$$

$$= (1 - \logit^{-1}(\theta_{1i} - b_{j1})) \times \frac{\exp(\sum_{l=0}^{2} (\theta_{2i} - b_{jl}))}{\sum_{l=0}^{2} \exp \sum_{m=0}^{l} (\theta_{2i} - b_{jm})} \cdot \frac{\exp(\theta_{3i})}{\sum_{l=0}^{2} \exp \sum_{m=0}^{l} (\theta_{3i} - b_{3jm})}.$$  

(3.8)

The probability of observing each of the possible inconclusive reasons (‘close’, ‘insufficient’, or ‘no overlap’) is made up of a Rasch probability and two partial credit model probabilities:

$$P(Y_{ij} = \text{Close}) = P(Y_{1ij}^* = 0) \times P(Y_{2ij} = 2 | Y_{1ij} = 0) \times P(Y_{3ij} = 1 | Y_{2ij} = 2)$$

$$= (1 - \logit^{-1}(\theta_{1i} - b_{j1})) \times \frac{\exp(\sum_{l=0}^{1} (\theta_{2i} - b_{2jl}))}{\sum_{l=0}^{2} \exp \sum_{m=0}^{l} (\theta_{2i} - b_{jm})} \times \frac{\exp(\theta_{3i})}{\sum_{l=0}^{2} \exp \sum_{m=0}^{l} (\theta_{3i} - b_{3jm})}.$$  

(3.11)

$$P(Y_{ij} = \text{Insufficient}) = P(Y_{1ij}^* = 0) \times P(Y_{2ij} = 2 | Y_{1ij} = 0) \times P(Y_{3ij} = 2 | Y_{2ij} = 2)$$

$$= (1 - \logit^{-1}(\theta_{1i} - b_{j1})) \times \frac{\exp(\sum_{l=0}^{1} (\theta_{2i} - b_{2jl}))}{\sum_{l=0}^{2} \exp \sum_{m=0}^{l} (\theta_{2i} - b_{jm})} \times \frac{\exp(\sum_{l=0}^{1} (\theta_{3i} - b_{3jl}))}{\sum_{l=0}^{2} \exp \sum_{m=0}^{l} (\theta_{3i} - b_{3jm})},$$  

(3.13)
and

\[
P(Y_{ij} = \text{No Overlap}) = P(Y_{1ij}^* = 0) \times P(Y_{2ij} = 2|Y_{1ij} = 0) \times P(Y_{3ij} = 3|Y_{2ij} = 2)
\]
\[= (1 - \logit^{-1}(\theta_{1i} - b_{1j})) \times \frac{\exp(\sum_{l=0}^{1} (\theta_{2i} - b_{2jl}))}{\sum_{l=0}^{2} \exp(\sum_{m=0}^{l} (\theta_{2i} - b_{2jm}))} \times \frac{\exp(\sum_{l=0}^{2} (\theta_{3i} - b_{3jl}))}{\sum_{l=0}^{2} \exp(\sum_{m=0}^{l} (\theta_{3i} - b_{3jm}))}.
\]

Furthermore, an item-explanatory variable \((X_j)\) for each item was included, where \(X_j = 1\) if the latent and reference print came from the same source (i.e. a true match) and \(X_j = 0\) if the latent and reference print came from different sources (i.e. a true non-match). Then,

\[
b_{jk} = \beta_{0k} + \beta_{1k}X_j + \epsilon_{jk},
\]

where \(b_{jk}\) are the item parameters and \(\beta_{0k}, \beta_{1k}\) are linear regression coefficients. In the IRT literature, this is known as the Linear Logistic Test Model (Fischer, 1973) with random item effects (Janssen et al., 2004). This allows for the means of item parameters to differ depending on whether the pair of prints is a true match or not. \(\epsilon_{jk}\) is also included to allow for variation in the item parameters within same-source (and different-source) pairs. Since pairs of prints are likely to have additional characteristics that impact their difficulty (e.g. image quality, number of features presents), not including \(\epsilon_{jk}\) in the model would be a very strong restriction, as all same-source pairs would have identical item parameters, as would all different-source pairs.

### 3.4.2 Binary decision process tree

In addition to the IRTree model representing the observed responses directly, I also constructed an IRTree based on one possible internal decision process. The first node, \(Y_{1i}^*\), represents the same latent value decision as the polytomous response tree. The second node, \(Y_{2i}^*\) represents whether there is sufficient information in the (reference, latent) pair to make a further decision. \(Y_{3i}^*\) represents whether the pair of prints is more likely to be a match or a non-match, and \(Y_{4i}^*\) and \(Y_{5i}^*\) represent whether this
determination is conclusive (individualization and exclusion, respectively) or inconclusive (close and no overlap, respectively). The tree structure is represented in Figure 3.5. The binary decision process tree thus separates decisions into both (a) distinguishing between matches and non-matches ($Y_i^*$) and (b) examiner “willingness to respond” ($Y_1^*, Y_2^*, Y_4^*, Y_5^*$).

Then, since each internal node is a binary split, a Rasch model can be used to parameterize each branch in the tree. That is, $P(Y_{kij} = 1) = \text{logit}^{-1}(\theta_{ki} - b_{kj})$, where $i$ indexes participants, $j$ indexes items, and $k$ indexes internal nodes. The full tree model is defined below:

$$P(Y_{ij} = \text{No Value}) = P(Y_{1ij}^* = 1)$$  \hspace{1cm} (3.18)

$$= \text{logit}^{-1}(\theta_{1i} - b_{1j})$$  \hspace{1cm} (3.19)

$$P(Y_{ij} = \text{Indiv.}) = P(Y_{1ij}^* = 0) \times P(Y_{2ij}^* = 0) \times P(Y_{3ij}^* = 1) \times P(Y_{4ij}^* = 1)$$  \hspace{1cm} (3.20)

$$= (1 - \text{logit}^{-1}(\theta_{1i} - b_{1j})) \times (1 - \text{logit}^{-1}(\theta_{2i} - b_{2j})) \times \text{logit}^{-1}(\theta_{3i} - b_{3j}) \times \text{logit}^{-1}(\theta_{4i} - b_{4j})$$  \hspace{1cm} (3.21)
\[ P(Y_{ij} = \text{Close}) = P(Y^*_{1ij} = 0) \times P(Y^*_{2ij} = 0) \times P(Y^*_{3ij} = 1) \times P(Y^*_{4ij} = 0) \]  
\[ = (1 - \text{logit}^{-1}(\theta_{1i} - b_{1j}))(1 - \text{logit}^{-1}(\theta_{2i} - b_{2j}))(\text{logit}^{-1}(\theta_{3i} - b_{3j})) \times (1 - \text{logit}^{-1}(\theta_{4i} - b_{4j})) \]  
\[ (3.22) \]

\[ P(Y_{ij} = \text{Insufficient}) = P(Y^*_{1ij} = 0) \times P(Y^*_{2ij} = 1) \]  
\[ = (1 - \text{logit}^{-1}(\theta_{1i} - b_{1j})) \times \text{logit}^{-1}(\theta_{2i} - b_{2j}) \]  
\[ (3.23) \]

\[ P(Y_{ij} = \text{No Ov.}) = P(Y^*_{1ij} = 0) \times P(Y^*_{2ij} = 0) \times P(Y^*_{3ij} = 0) \times P(Y^*_{5ij} = 0) \]  
\[ = (1 - \text{logit}^{-1}(\theta_{1i} - b_{1j})) \times (1 - \text{logit}^{-1}(\theta_{2i} - b_{2j})) \times (1 - \text{logit}^{-1}(\theta_{3i} - b_{3j})) \times (1 - \text{logit}^{-1}(\theta_{5i} - b_{5j})) \]  
\[ (3.24) \]

\[ P(Y_{ij} = \text{Excl.}) = P(Y^*_{1ij} = 0) \times P(Y^*_{2ij} = 0) \times P(Y^*_{3ij} = 0) \times P(Y^*_{5ij} = 1) \]  
\[ = (1 - \text{logit}^{-1}(\theta_{1i} - b_{1j})) \times (1 - \text{logit}^{-1}(\theta_{2i} - b_{2j})) \times (1 - \text{logit}^{-1}(\theta_{3i} - b_{3j})) \times \text{logit}^{-1}(\theta_{5i} - b_{5j}) \]  
\[ (3.25) \]

with

\[ b_{kj} = \beta_{0k} + \beta_{1k}X_j + \epsilon_{kj}, \]  
\[ (3.30) \]

as in the observed response tree above.

This binary decision tree model has a few advantages over the observed response tree. First, it represents an idealized process of reporting decisions, where each sub-decision is clearly defined and separable from the rest of the process. This leads to interpretable parameters that may generalize neatly to reporting schemes that impose this ordering explicitly. Interpretable parameters are especially
desirable in this setting, as they could provide clear recommendations for targeted training. Second, each of the branches have been constructed to be binary. This not only helps with interpretability, but removes the ordered responses assumption that is present in polytomous trees.

3.4.3 Simplified scored tree

Finally, to assess whether the more complicated tree models above provide added value over a simpler model, a simplified scored tree model was also fit to the data, where the source decision is scored as correct or incorrect. That is, \( Y^*_1 \) is the same value assessment as in the above models, and \( Y^*_2 \) represents the conclusive source decisions (i.e. individualization or exclusion) that is simply scored as correct or incorrect. Although scoring inconclusive responses as correct or incorrect is an issue in itself (see Chapter 2 and Luby (2019) for discussion), for the purposes of this work inconclusive responses were scored as incorrect.

As in the binary process tree, each node is modeled with Rasch probabilities:

\[
P(Y_{ij} = \text{No Value}) = P(Y^*_{1ij} = 1) = \logit^{-1}(\theta_{1i} - b_{1j})
\]

\[
P(Y_{ij} = \text{Correct}) = P(Y^*_{1ij} = 0) \times P(Y^*_{2ij} = 1|Y^*_{1ij} = 0) = (1 - \logit^{-1}(\theta_{1i} - b_{1j})) \times \logit^{-1}(\theta_{2i} - b_{2j})
\]
\[ P(Y_{ij} = \text{Incorrect}) = P(Y^*_{1ij} = 0) \times P(Y^*_{2ij} = 0|Y^*_{1ij} = 0) \]
\[ = \logit^{-1}(\theta_{1i} - b_{1j}) \times (1 - \logit^{-1}(\theta_{2i} - b_{2j})) \]  

(3.35)  (3.36)

### 3.4.4 Prior Distributions

We fit the three models under the Bayesian framework with Stan in R (Stan Development Team, 2018a; R Core Team, 2013). For each model the following prior distributions were assigned:

\[ \theta \sim MVN(0, \sigma_\theta L_\theta L_\theta') \]
\[ b \sim MVN(X\beta, \sigma_b L_b L_b') \]
\[ L_\theta \sim LKJ(4) \]
\[ L_b \sim LKJ(4) \]
\[ \sigma_\theta \sim \text{Half-Cauchy}(0, 2.5) \]
\[ \sigma_b \sim \text{Half-Cauchy}(0, 2.5) \]
\[ \beta_k \sim N(0, 5) \]

Multivariate normal distributions for \( \theta \) and \( b \) were chosen to estimate covariance between latent traits explicitly. The Stan modeling language does not rely on conjugacy, so the Cholesky factorizations (\( L_\theta \) and \( L_b \)) are modeled instead of the covariance matrices for computational efficiency. The recommended priors (Stan Development Team, 2018b) for \( L \) and \( \sigma \) were used: an LKJ prior (Lewandowski et al., 2009) (LKJ = last initials of authors) with shape parameter 4, which results in correlation matrices which mildly concentrate around the identity matrix (LKJ(1) results in uniformly sampled correlation matrices), and half-Cauchy priors on \( \sigma_b \) and \( \sigma_\theta \) to weakly inform the correlation between latent traits. \( X \) are the item covariates (as in Equation 3.17), and \( N(0, 5) \) priors were assigned to the linear regression coefficients (\( \beta_k \)).
3.5 Results

This section first addresses model fit. Since the binary process tree and the ordered response tree use the same response variable, their respective model fits are compared. The simplified-scored tree uses a different response variable (since the source decision is scored) and is therefore not comparable to the other two models using standard model evaluation techniques. Next, the parameter estimates from the three models are compared to one another and their usefulness assessed. Finally, the questions raised about inconclusive responses in Sections 2.3.2 and 3.3 are addressed through a model-based way to distinguish erroneous inconclusives from correct inconclusives.

3.5.1 Model Fit Comparison

Since both the ordered response tree and the binary decision process tree retain unique labels for all six possible responses (individualization, exclusion, inconclusive-close, inconclusive-insufficient, inconclusive-no overlap, no value), the prediction error can be used to compare the fit of the two models. That is, for each observation, the probability of observing each of the six possible responses for the relevant (examiner, item) pair was calculated. The predicted value is the response with the highest estimated probability. The simplified-scored tree, however, transforms the observations to either ‘No Value’, ‘Correct’, or ‘Incorrect’, which makes the model fit incomparable to the other two models. We will return to the simplified-scored tree in the next section.

The in-sample prediction error of the response-based tree is 0.22 while the prediction error for the binary process tree is 0.19, suggesting that the binary process tree does a slightly better job of explaining responses than the response-based tree.

Although the binary process tree and the observed response tree use the same outcome variable, the binary process tree estimates more parameters than the observed response tree (Table 3.2), and this difference must be accounted for when comparing prediction error.
TABLE 3.2: Number of parameters in each model

<table>
<thead>
<tr>
<th></th>
<th># of parameters per examiner</th>
<th># of parameters per item</th>
<th>Total # of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Response</td>
<td>3</td>
<td>5</td>
<td>4,277</td>
</tr>
<tr>
<td>Binary Decision Process</td>
<td>5</td>
<td>5</td>
<td>4,565</td>
</tr>
<tr>
<td>Simplified-scored</td>
<td>2</td>
<td>2</td>
<td>1,826</td>
</tr>
</tbody>
</table>

WAIC (Watanabe, 2010) as defined in Vehtari et al. (2017) was used to compare the models to one another. WAIC is a fully Bayesian information criterion and closely approximates Bayesian cross-validation, without the need to re-fit on different training sets.

\[
\text{WAIC} = -2 \times \left( \hat{l}_{pd} - \sum_{n=1}^{N} \sum_{s=1}^{S} V \log(p(y_n|\theta^s, b^s)) \right)
\]

where \( n \) indexes each participant \( \times \) item pair, \( V \) represents the sample variance, \( \theta^s \) and \( b^s \) are the parameter estimates from the \( s^{th} \) posterior draw, and \( \hat{l}_{pd} = \sum_{n=1}^{N} \log \left( \frac{1}{S} \sum_{s=1}^{S} p(y_n|\theta^s, b^s) \right) \).

TABLE 3.3: WAIC (with standard errors) for the Observed Response tree and Binary Response tree models, demonstrating that the Binary Response tree leads to a better predictive performance.

<table>
<thead>
<tr>
<th>Model</th>
<th>WAIC</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Response</td>
<td>29736</td>
<td>348</td>
</tr>
<tr>
<td>Binary Process</td>
<td>21985</td>
<td>238</td>
</tr>
</tbody>
</table>

Finally, the fit of the two models was evaluated using a posterior predictive check. Within each posterior draw, the outcome (i.e. ‘no value’, ‘individualization’, ‘close’, etc.) was predicted for each participant \( \times \) item pair based on their associated parameters. Each participant was then summarized by six predicted scores per posterior draw: (1) The number of ‘No Value’ observations, (2) the number of predicted ‘individualization’ observations, (3) the number of predicted ‘close’ observations, (4) the number of predicted ‘insufficient’ observations, (5) the number of predicted ‘no overlap’ observations, and (6) the number of predicted ‘exclusion’ observations. The 95% posterior predictive intervals for each score can be calculated by taking the 2.5 and 97.5 percentiles. These posterior predictive intervals are plotted in Figure 3.7 against the actual observed score for each person. If a model was predicting perfectly, all bars would fall along the diagonal line \( y = x \) (denoted with a dotted line).
As expected, the models perform nearly identically for the ‘no value’ outcome. They also perform similarly for the ‘individualization’ and ‘close’ outcomes. The models begin to diverge for the ‘insufficient’, ‘no overlap’, and ‘exclusion’ outcomes. The observed response model severely under-predicts insufficient responses as the number of observed insufficient responses grows. The same model both over and under-predicts the ‘no overlap’ outcome. It also over-predicts the number of ‘exclusion’ outcomes when there are few exclusions observed. Although the binary decision process model does not perform perfectly, for example, slightly over-predicting individualizations and exclusions when there were few observed, this check provides evidence that the binary decision process model is outperforming the observed response model in predicting outcomes at the participant level.

### 3.5.2 Parameter estimates

A benefit of using a Bayesian estimation method is that an entire posterior sample for each parameter is obtained instead of a single point estimate. For instance, if we consider the first split in the binary process tree (‘No Value’ or ‘Has Value’), the $j^{th}$ item is associated with a parameter, $b_{1j}$ (i.e Eq. 17), that governs the item effect for that decision. Instead of a single numeric estimate for $b_{1j}$, the entire posterior distribution for $b_{1j}$ is estimated.
There are 744 questions in the study, however, and looking at each of the posterior densities individually, for each of 5 parameters, would not lead to any clear insights. It is perhaps more useful to assess overall trends for the items and examiners using the point estimates. The posterior median for each examiner and item were calculated, and the distribution of examiner parameters (Figures 3.8, 3.11, 3.14) and item parameters (Figures 3.9, 3.12, 3.15) are displayed as a whole.

In all three models, the item parameters are generally more extreme than the person parameters corresponding to the same decision. This suggests that many of the responses are governed by item effects, rather than examiner tendencies. There is, however, notable variation in both person and item parameters across all three models.

In the binary decision response tree, the greatest variation in person parameters occurs in θ₁ (‘no value’ tendency), θ₄ (conclusive tendency in matches) and θ₅ (conclusive tendency in non-matches). Item parameters are most extreme in b₁ (tendency towards has value) and b₄ (inconclusive tendency
in matches). There are therefore some items that all examiners would be expected to agree on those decisions (has value vs no value and individualization vs close).

In the ordered response tree, there is very little variation in $\theta_2$ (exclusion tendency) and $\theta_3$ (No overlap tendency) as compared to $\theta_1$ (no value tendency). Compared to the binary decision process tree, this model picks up very few differences among examiners in their source assessments or inconclusive reasoning.

In the simplified-scored tree, the distributions of no value tendency parameters ($\theta_1$ and $b_1$) and the parameters that describe correct/incorrect decisions ($\theta_2$ and $b_2$) are relatively similar to one another, without many of the trends observed in the other models.

Notice that there is some bimodality in the item parameters (especially noticeable in the parameters for $Y^*_3$ in the Observed Response tree (in yellow), and the sufficient tendency in the binary decision process tree), which is due to including an indicator covariate in the item parameters (Eq. 16) when the
pair is a true match. These bimodal distributions are then representative of two clusters of estimates: one for true matches and one for true non-matches.

Recall that $Y^*_1$ in each of the three IRTrees represents the same decision (‘no value’ vs ‘has value’) in each of the three models, so one would expect to see very similar estimates for the corresponding parameters. Figure 3.17 shows the $b_1$ and $\theta_1$ estimates for the observed response tree and the simplified-scored tree compared to the binary decision process tree. There is, as expected, a very strong linear relationship between the parameters in each model. The variation at large values of $b_1$ is due to those items being rated as ‘has value’ by every examiner who saw the question. Since there is no variation in examiner response, there is no information in the data to precisely locate $b_1$ for those items. This variation across models for large $b_1$ values does not have any practical significance, since each will correctly predict that these items will be found to have value.

Although the simplified scored model does not contain as much detail about examiner decisions as the other two IRTree models, $\theta_2$ in the simplified-scored model is the only parameter in any of the models that is a traditional measure of ‘proficiency’, and thus provides a useful comparison metric for the other parameters. Figure 3.18 shows the relationship between this proficiency estimate with the person parameters estimated in the other models. There is a stronger linear relationship between proficiency ($\theta_2$ in the simplified-scored model) and the estimated person effects in the binary decision process tree than in the observed response tree, suggesting that proficiency is at least encoded in some
of the parameters in the binary decision process tree. This is further evidence that the binary decision response tree provides a more meaningful way to analyze the sequential responses in the black box study than treating the observed responses as ordered categories.

### 3.5.3 Question Ambiguity vs Examiner Indecision

As previously discussed in Chapter 2, inconclusive responses are quite prevalent in the data. Inconclusive responses can happen for two reasons: 1) the latent print quality was too poor to conclude anything or 2) the examiner was unsure of his/her response. Inconclusive responses that are due to poor quality items should not be penalized, while the inconclusive responses due to examiner indecision should be treated as errors. One would like to distinguish between the two, but do not have any additional information about the items (e.g. image quality, number of minutiae, etc.) to inform this distinction.

One strategy for determining whether an observed inconclusive is due to examiner indecision is to calculate the probability of a conclusive response under the model, and if that probability is sufficiently high, to label the inconclusive as due to examiner indecision rather than print quality.

For instance, Examiner 55 (ID = A2009) decided Question 556 (ID = N057413) was a ‘Close’ inconclusive, but Question 556 is a true non-match. Using the estimates for $\theta_{55,k}$ and $b_{556,k}$ under the binary decision process model (where $k = 1, \ldots, 5$ and indexes each split in the tree), the probability of observing each response for this observation can be calculated. Using point estimates of $\theta_{55,k}$ and $b_{556,k}$,
this results in \(P(\text{No Value}) < 0.005, P(\text{Individualization}) < 0.005, P(\text{Close}) = 0.20, P(\text{Insufficient}) < 0.005,\)
\(P(\text{No Overlap}) = 0.01\) and \(P(\text{Exclusion}) = 0.78.\) According to the model, the most likely outcome for this response is an exclusion. Since an inconclusive was observed instead, this response might be flagged as being due to examiner indecision.

We calculate \(P(\text{Exclusion}), P(\text{Individualization})\) and \(P(\text{Conclusive})\) (where \(P(\text{Conclusive}) = P(\text{Exclusion}) + P(\text{Individualization})\)) for each examiner \(\times\) item observation in the data. Each quantity is then separated into two groups: one group of observed conclusives and one of observed inconclusives. There should be clear separation in \(P(\text{Conclusive})\) between the two groups. Observed conclusives should have high \(P(\text{Conclusive})\) and observed inconclusives should have low \(P(\text{Conclusive})\).

Histograms of each examiner \(\times\) item pair are shown in Figure 3.19 binned by the calculated \(P(\text{Conclusive})\) under the binary decision process model. The observed inconclusive responses are shown in green and and the observed conclusive responses are shown in purple. As expected, the inconclusive responses largely have low values of \(P(\text{Conclusive})\) while conclusive responses have large \(P(\text{Conclusive})\). The observed inconclusives which have a high \(P(\text{Conclusive})\) should have been conclusive responses (according to the model), and are therefore likely due to examiner indecision.
There is significant overlap for the middle-range values of P(Conclusive), so we must determine the point at which P(Conclusive) is “high enough” to classify an inconclusive response as an error.

A potential cutoff point at $c = 0.5$ is illustrated using a dashed line, but a more effective cutoff point may be either higher or lower. The truly optimal cutoff cannot be determined because we do not have access to the item content (e.g. the ‘true’ quality of the print) nor to the loss function used by law enforcement (e.g. a false exclusion may be a more egregious error than an inconclusive error). Instead, the predictive performance of different cutoff points can be determined by determining the percent of True Matches that were predicted to be individualizations and the percent of True Non-matches that were predicted to be exclusions. In other words, a classifier can be constructed for the observed inconclusive responses using their predicted probabilities as a decision rule. For instance, using $P(Conclusive) > c$ with $c = 0.5$ as a cutoff for classifying erroneous inconclusives leads to the confusion matrix in Table 3.4.

---

**Table 3.4: Confusion matrix for inconclusive responses using $P(Conclusive) > .5$ as a decision rule.**

<table>
<thead>
<tr>
<th></th>
<th>Predicted Conclusive</th>
<th>Predicted Other</th>
<th>Not Flagged</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Matches</td>
<td>422</td>
<td>82</td>
<td>3371</td>
</tr>
<tr>
<td>True Non-Matches</td>
<td>299</td>
<td>0</td>
<td>733</td>
</tr>
</tbody>
</table>

---

**Figure 3.19:** Histograms of conclusive examiner × item observations (purple) and inconclusive examiner × item observations (green), binned by estimated $P(Conclusive)$. For example, $P(Conclusive) > 0.5$ (dotted line) could be used as a decision rule to classify observed inconclusives as erroneous.
The performance of various cutoff points can then be compared using an ROC-type plot (Figure 3.20). Instead of the traditional ROC quantities, the Correct ID Prediction Rate and Correct Exclusion Prediction Rate are used, where

\[
\text{Correct ID Prediction Rate} = \frac{\text{True Matches that were Predicted IDs with } P(\text{Conc}) > c}{\text{All True-match observations with } P(\text{Conc}) > c} \quad (3.38)
\]

and

\[
\text{Correct Exclusion Prediction Rate} = \frac{\text{True non-matches predicted exclusion with } P(\text{Conc}) > c}{\text{All True-non-match observations with } P(\text{Conc}) > c}. \quad (3.39)
\]

For Table 3.4 when \( c = 0.5 \), the correct ID prediction rate is thus \( \frac{422}{422+82} = 0.837 \) and the correct exclusion prediction rate is \( \frac{299}{299+0} = 1 \). When \( c = 0 \), the Correct Individualization Rate and the Correct Exclusion Rate are low because all of the inconclusive observations are included in the denominators. When \( c = 1 \) the Correct Individualization Rate and Correct Exclusion Rate are both one because none of the inconclusive observations are included in the calculations. The complete curve, using \( 0 \leq c \leq 1 \) in increments of 0.1, is shown in Figure 3.20. The Correct Individualization Rate and Correct Exclusion Rate rapidly increase as \( c \) increases from zero to 0.5, and the number of inconclusives classified as errors rapidly decreases, but there is not a noticeable difference in performance as \( c \) increases past 0.5.

Using \( P(\text{Conclusive}) \) to determine which inconclusives were ‘mistakes’ is not ideal, however. It is possible for an observation to have \( P(\text{Individualization}) = 0.30, P(\text{Exclusion}) = 0.30, \) and \( P(\text{Insufficient}) = 0.40, \) for instance. This would result in \( P(\text{Conclusive}) = 0.6 \), even though the model found ‘Insufficient’ to be the most likely outcome. A better approach may be to look at \( P(\text{Exclusion}) \) and \( P(\text{Individualization}) \) separately, and only classify inconclusives as an error if either of those probabilities are sufficiently high. Histograms of these probabilities are shown in Figure 3.21 split by whether the item was a true match or a true non-match.
FIGURE 3.20: Performance of procedure to classify erroneous inconclusives at various cutoff points \( c \). As \( c \) increases from 0 to 0.5, correct individualization rate and correct exclusion rate increases and the number of inconclusives classified as errors decreases. There is no noticeable difference in performance when \( c > 0.5 \). Correct Individualization Rate and Correct Exclusion Rate are as defined in Equations 3.38 and 3.39, respectively.

FIGURE 3.21: Estimated probability of observing an individualization (for matches) or an exclusion (for true non-matches) for the observed inconclusive responses.

Similarly, a ‘classifier’ for the inconclusive responses can be constructed using \( P(\text{Individualization}) > c \) or \( P(\text{Exclusion}) > c \) as a decision rule (rather than \( P(\text{Conclusive}) > c \)). This procedure leads to the curve in Figure 3.22. Using \( P(\text{Individualization}) \) and \( P(\text{Exclusion}) \) to classify inconclusives as erroneous leads to more effective classifications than \( P(\text{Conclusive}) \), particularly at low cutoff points.

Using probabilities calculated from the IRTree model estimates provides a way to assess the observed decisions for an examiner \( \times \) item pair in light of other decisions that examiner made, and how other examiners evaluated that item. Inconclusives that are ‘expected’ under the model can then be determined, along with which examiners often make decisions that are consistent with the model-based
predictions. For example, an examiner whose decisions often match the model-based predictions may be more proficient in recognizing when there is sufficient evidence to make a conclusive decision than an examiner whose decisions do not match the model-based predictions. With some adjustments, this type of classifier based on predicted probabilities has the potential to provide a model-based scoring of examiner decisions.

3.6 Discussion

We originally set out to use IRTrees to model differences in each step of ACE-V, the procedural standard for latent prints. ACE-V, however, does not consist of well-defined steps and so it is hard to construct a ‘true’ IRTree for the decision-making process. Different examiners may use different trees, but there is no way to determine or model those differences with existing data. Additional data should be collected to explicitly measure what decisions different examiners make in each step of the ACE-V process. A different reporting standard in which each decision-making step is well-defined and clearly separable from other decisions would better allow for better measurement with an approach such as IRTrees.

Instead of modeling the ACE-V decision process, we have modeled different sub-decisions that are distinguishable in the Black Box data. We have demonstrated how IRTrees may be constructed
in various ways (Section 3.4), provided a model comparison framework using both model selection methods and parameter comparisons (Section 3.5.1 and 3.5.2), and developed a model-based method to identify unexpected inconclusive responses (Section 3.5.3).

Three IRTree models were constructed to analyze the FBI “black box” study: one based on the observed response process, one based on a binary decision response process, and one in which the responses were scored. Although the scored response tree is not directly comparable to the other two, since the response variable is changed, there is evidence that the binary decision response tree provides a better explanation of the data than the observed response tree. We found that the binary decision process tree performed better than the observed response tree according to both WAIC and a posterior predictive check. Furthermore, the participant parameters from the binary decision response tree were correlated with examiner proficiency from the scored tree, while the participant parameters from the observed response tree were not.

We have also provided a model-based way to evaluate whether inconclusive responses were erroneous or not. By incorporating estimates about the item and examiner, one can evaluate whether inconclusive responses are due to insufficient information in the question or examiner indecision. Although more work is required to develop this approach fully, a model-based scoring system could be useful for additional applications where the data is not scored and there is no clear way to do so.

Finally, a bipolar scale for evaluating whether two prints came from the same source or different sources was provided. Although the examiners in the black box study were not provided with this scale, it does provide insight into examiner behavior and item effects, particularly for the inconclusive responses. This scale can be discretized or continuous and allows for the reporting scheme from the black box study as well as a likelihood-ratio based method of reporting.
3.7 Conclusions

How can collapsing multiple responses into one response (and therefore losing information) be avoided?

The IRTrees framework provides a useful way to model data where responses are made up of sequential decisions. Using a tree structure, each sub-decision in a response is modeled with an IRT model, allowing for separation of decision-maker and item effects at each step in the decision-making process. These sub-decisions are not required to represent a choice between a correct option and an incorrect option, which is especially useful for analyzing responses that are not scored.

Can the sequential nature of fingerprint analysis be preserved?

The structure of the IRTree model can retain the sequential nature of fingerprint decisions, since the IRT model at each branch is conditional on the preceding branches. Note that throughout in this chapter, only the observed response tree and simplified scored tree truly retain the sequential responses due to the recording of the data.

How does modeling differences in decision-making at each step in the process compare to traditional measures of proficiency?

As discussed above, the participant parameters from the binary decision response tree were correlated with examiner proficiency from the scored tree. Modeling the sequential decision explicitly does therefore not contain any less information that is available in a more standard IRT approach.
Chapter 4

Determining Expected Answers from Responses Alone

In forensic decision-making studies or proficiency tests, test administrators generally know whether the evidence came from the same-source or different sources. It is often harder to determine beforehand whether there is sufficient information in the evidence for most examiners to come to a conclusive decision. The purpose of this chapter is to use test responses, without additional information about the items, to generate an expected ‘answer key’ that does not depend on individual examiner tendencies.

The specific questions we would like to answer are:

1. How can response data for which there is an unknown correct answer be modeled within the IRT framework?

2. When participants are allowed to specify an uncertain response, how can it be determined if that uncertainty is justified or a mistake?
4.1 Introduction

As evidenced in Chapter 2, traditional item response models are not entirely appropriate for the FBI Black Box data, in large part because the dataset does not contain keyed correct responses. One of four broad responses is expected for each item: No Value, Inconclusive, Individualization, or Exclusion. Since the FBI has provided information about which pairs of prints came from the same source and which came from different sources, it is known whether individualization or exclusion responses are correct or incorrect. There is no obvious way, however, to determine when a no value or an inconclusive answer is expected.

Chapter 2 demonstrated the difficulty in scoring inconclusives, and Chapter 3 demonstrated the importance of distinguishing between inconclusive and conclusive responses. In practice, inconclusive responses that should have been individualizations can lead to guilty perpetrators not being prosecuted, while inconclusive responses that should have been exclusions lead to an inefficient use of resources through the additional investigation of innocent suspects. On the other hand, individualizations that should have been inconclusives can lead to innocent suspects being convicted, while exclusions that should have been inconclusives can lead to guilty perpetrators going free.

It is also important to distinguish between items for which a no value is expected and items for which an inconclusive is expected. A ‘true’ no value is a latent print of very poor quality, where an individualization or exclusion cannot be made regardless of the reference print. A ‘true’ inconclusive, on the other hand, consists of a latent print that may be very low quality, but does contain some information and therefore may be appropriate for an individualization or exclusion if an appropriate reference print is available (Ulery et al., 2011 SI 1.5). In practice, latent prints that are incorrectly rated as ‘no value’ may lead to evidence not being used to its full potential, with investigative leads or convictions being missed. Latent prints that have no value, but are not rated as such, lead to a labor and time intensive analysis process with a negligible chance of success.

From a training perspective, it is therefore important for examiners to have feedback not only when they make false identifications or exclusions, but also when false ‘no value’ or ‘inconclusive’ decisions
are made. In order to provide such feedback, it must be determined when ‘no value’ or ‘inconclusive’
decisions are expected and when ‘individualization’ or ‘exclusion’ decisions are expected. In the Black
Box study, no additional information about the items has been provided to generate an answer key. A
method for determining expected answers from the responses alone is therefore needed.

Chapter 3 used the IRTrees framework to account for the FBI black box responses arising through a
series of sequential decisions. Although the primary interest of Chapter 3 was understanding differences
in item and examiner tendencies at each step in the decision-making process, we coincidentally avoided
scoring the no value and inconclusive responses by modeling the sequential decision process explicitly.
Using the estimated item parameters, the most likely outcome for each item can be determined
independently of participant tendencies. Item Response Trees thus represent a way to determine the
expected responses, or answer key, for each item.

There are also existing methods for ‘IRT without an answer key’, namely the cultural consensus
theory (CCT) approach (Batchelder and Romney, 1988; Oravecz et al., 2014). CCT was designed for
situations when a group of respondents shares some knowledge or beliefs in a domain area which is
unknown to the researcher or administrator. CCT then estimates the expected answers to the questions
provided to the group.

This chapter uses both CCT and IRTrees to find expected responses (of the four outcomes discussed
above) for the FBI Black Box data based on responses alone. Section 4.2 provides a brief overview of
relevant CCT and IRTree literature. Section 4.3 discusses a simplified version of the IRTree model from
Chapter 3, introduces an existing CCT model for ordered responses, and proposes two adapted CCT
models for this setting. Although the CCT and IRTree models are motivated and parameterized quite
differently, in this setting they produce similar answer keys. Section 4.4 derives conditional sufficient
statistics, or shows that nontrivial conditional sufficient statistics do not exist, for each of the models
considered, which provides a theoretical justification for the similarities between the two models.
Section 4.5 summarizes the results of these models as applied to the FBI black box study, and Section 4.6
discusses broad conclusions and recommendations.
4.2 Literature

Cultural Consensus Theory (CCT) is one approach for “scoring without an answer key” (Batchelder and Romney, 1988) that was initially developed to identify culturally-shared information in a particular group. It is similar to IRT in that individual-specific parameters are used to estimate the probability of a given response to a stimuli, however the ‘correct’ answers are not known beforehand. CCT models have been developed for binary “answer keys” (Oravecz et al., 2015) and continuous scales (Anders et al., 2014), as well as for cases when different populations have different answer keys (Anders and Batchelder, 2012).

This chapter primarily compares the Latent Truth Rater Model (LTRM), a CCT model for ordinal categorical responses (Anders and Batchelder, 2015) to an IRTree-based approach. The LTRM assumes that each item is located on a latent continuum, and each participant’s assessment of that location varies based on person and item effects. It therefore estimates “culturally shared” item locations and category thresholds, as well as response bias tendencies and competencies for each participant. The model specifications are discussed in more detail in Section 4.3.1.

Item Response Trees (De Boeck and Partchev, 2012) are relatively new to the IRT literature and are the second methodological focus of this chapter. They are often used to model differences in response behavior (e.g. fast/slow responses or extreme response styles) in tests or surveys (Sinayev and Peters, 2015; Plieninger and Meiser, 2014; Böckenholt, 2017; Debeer et al., 2017). IRTrees have also been proposed to model behavior in non-traditional IRT applications. Cursio et al. (2019) illustrate the use of an IRTree with a single latent trait to model high school students’ mood. Lopez-Sepulcre et al. (2015) recommended IRTrees to model sequential behavioral data in animals. IRTrees have not, to my knowledge, been used to determine an answer key for a set of response data.

A further approach to modeling non-standard response data was developed for aptitude tests, particularly within the Office of Naval Research (Sympson, 1985, 1981), with a different parameterization of the model in Thissen and Steinberg (1984). This approach seeks to preserve the information present
in response options (including non-response) by estimating many parameters for each item using a
generalized additive model-like approach (Sympson [1983]).

4.3 Model Formulation

4.3.1 Polytomous Ordered Response CCT (Latent Truth Rater Model)

In order to fit the CCT model, the responses must be ordered, which is not a requirement of the IRTree
approach. Furthermore, the ‘no value’ responses should be included, but they do not explicitly fit on the
exclusion - individualization scale introduced in Chapter 3 (Reproduced in Figure 4.1 below).

![Figure 4.1: FBI black box responses on an individualization/exclusion scale, as used in Chapter 3](image)

Although the individualization/exclusion scale in Figure 4.1 could be used to generate an answer
key for the source decisions (i.e. individualization, exclusion, or inconclusive), it would not be possible
to determine an answer key for the latent value decisions (i.e. no value vs has value). Since the latent
value decision assesses a single latent print, instead of a pair of prints, there is not an obvious location
on the individualization/exclusion scale that corresponds to a no value decision, since both exclusions
and individualizations are inherently based on a comparison of two prints. The FBI has provided
whether the prints originated from the same source or different sources, and so a ‘conclusiveness’ scale
(Figure 4.2) can be used instead. This scale does not distinguish between same source/different source
prints, but does allow for the inclusion of no value responses on the scale. Using an answer key from this
scale, alongside the same-source/different-source information provided by the FBI, provides a complete
picture of what the expected answers are (e.g. conclusive same-source pairs should be individualizations
and conclusive different-source pairs should be exclusions).
Let $Y_{ij}$ denote participant $i$’s response to item $j$. The LTRM is largely explained by $T_j$, the latent “answer key” for item $j$, and $\gamma_c$ ($c = 1, 2$), the category boundaries between ‘No Value’ and ‘Inconclusive’ and ‘Inconclusive’ and ‘Conclusive’, respectively. Each participant draws a latent appraisal of each item ($Z_{ij}$), which is assumed to follow a normal distribution with mean $T_j$ (the ‘true’ location of item $j$) and precision $\tau_{ij}$, which depends on both participant competency ($E_i$) and item difficulty ($\lambda_j$) (that is, $\tau_{ij} = \frac{E_i}{\lambda_j}$). If every participant uses the ‘true’ category boundaries, then if $Z_{ij} \leq \gamma_1$ then $Y_{ij} = ‘No Value’$, if $\gamma_1 \leq Z_{ij} \leq \gamma_2$ then $Y_{ij} = ‘Inconclusive’$, and if $Z_{ij} \geq \gamma_2$ then $Y_{ij} = ‘Conclusive’$. Individuals, however, might use a biased form of the category thresholds, and so individual category thresholds, $\delta_{i,c} = a_i \gamma_c + b_i$, are defined, where $a_i$ and $b_i$ are participant scale and shift biasing parameters, respectively. That is, $a_i$ shrinks or expands the category thresholds for participant $i$, and $b_i$ shifts the category thresholds to the left or right. The model is thus defined through the following:

$$P(Y_{ik} = \text{No Value}) = P(Z_{ik} \leq \delta_{i,1})$$

$$= P(T_k + \epsilon_{ik} \leq a_i \gamma_1 + b_i)$$

$$= F(a_i \gamma_1 + b_i)$$
\[ P(Y_{ik} = \text{Inconclusive}) = P(\delta_{i,1} < Z_{ik} \leq \delta_{i,2}) \]
\[ = P(a_i \gamma_1 + b_i \leq T_k + \epsilon_{ik} \leq a_i \gamma_2 + b_i) \]
\[ = F(a_i \gamma_2 + b_i) - F(a_i \gamma_1 + b_i) \]  

\[ P(Y_{ik} = \text{Conclusive}) = P(Z_{ik} > \delta_{i,2}) \]
\[ = P(T_k + \epsilon_{ik} > a_i \gamma_2 + b_i) \]
\[ = 1 - F(a_i \gamma_2 + b_i) \]  

where \( F(u) \) is the CDF of a normal variable with mean \( T_k \) and precision \( \tau_{ik} \):

\[ F(u) = \sqrt{\frac{\tau_{ik}}{2\pi}} \int_{-\infty}^{u} \exp\left(-\frac{1}{2} \tau_{ik} (y - T_k)^2\right) dy. \]  

A graphical representation of the LTRM category probabilities is shown in Figure 4.3 for a question with \( T_j = 0 \).

The likelihood of the data under the LTRM is then:
\[
L(Y|T, a, b, \gamma, E, \lambda) = \prod_I \prod_J \left[ F_{y_{ij}}(\delta_{i,x_{ij}}) - F_{y_{ij}}(\delta_{i,x_{ij}-1}) \right]
\]

where \(\delta_{i,0} = -\infty, \delta_{i,C} = \infty,\) and \(\delta_{i,c} = a_i \gamma_c + b_i.\)

### 4.3.2 Adapted LTRM as a Cumulative Logits Model

The original LTRM (Equation 4.11) is a cumulative-probits model, and is therefore more closely related to more standard IRT models than it might seem at first glance. Specifically, if (1) the latent appraisals \((Z_{ij})\) are modeled with a logistic instead of a normal distribution, (2) it is assumed that \(\tau_{ij} = \frac{E_i}{\lambda_j} = 1\) for all \(i, j,\) and (3) it is assumed \(a_i = 1\) for all \(i,\) then the model collapses into a more familiar cumulative logits IRT model:

\[
\log \frac{P(Y_{ij} \leq c)}{P(Y_{ij} > c)} = b_i - T_j + \gamma_c.
\]

This transformed model has the same form as the Graded Response Model \([\text{Samejima}, 1969]\) (See Chapter 2 for overview).

Relaxing the assumption that \(a_i = 1,\) a cumulative logits model with a scaling effect for each person on the item categories is obtained, which we call the cumulative-logits LTRM (C-LTRM):

\[
\log \frac{P(Y_{ij} \leq c)}{P(Y_{ij} > c)} = b_i - T_j + a_i \gamma_c.
\]

The likelihood for the data under Equation 4.13 is:

\[
L(Y|a, b, T, \gamma) = \prod_I \prod_J \left[ \frac{\exp(b_i - T_j + a_i \gamma_c)}{1 + \exp(b_i - T_j + a_i \gamma_c)} - \frac{\exp(b_i - T_j + a_i \gamma_{c-1})}{1 + \exp(b_i - T_j + a_i \gamma_{c-1})} \right]
\]

where \(\gamma_0 = -\infty\) and \(\gamma_C = \infty.\)
4.3.3 Adapted LTRM as an Adjacent Category Logits Model

Making the same assumptions as above, \( P(Y_{ij} = c) \) could instead be expressed using an adjacent-categories logit model:

\[
\log \frac{P(Y_{ij} = c)}{P(Y_{ij} = c - 1)} = b_i - T_j + \gamma_c
\]  

(4.15)

which takes the same form as the Rating Scale Model (Andrich, 1978) (See Chapter 2 for overview). The RSM has nice theoretical properties due to the separability of \( T_j \) and \( b_i \) in the likelihood, and re-casting the LTRM as an adjacent-categories model opens the possibility of more direct theoretical comparisons between models.

Relaxing the assumption that \( a_i = 1 \), a generalized adjacent-categories logit model with a scaling effect for each person on the item categories is obtained, which we call the adjacent-logits LTRM (A-LTRM):

\[
\log \frac{P(Y_{ij} = c)}{P(Y_{ij} = c - 1)} = b_i - T_j + a_i\gamma_c.
\]  

(4.16)

The likelihood is then:

\[
L(Y | a, b, T, \gamma) = \prod_I \prod_J \frac{\exp(b_i - T_j + a_i\gamma_c)}{1 + \exp(b_i - T_j + a_i\gamma_c)}.
\]  

(4.17)

4.3.4 IRTree for answer key generation

I have constructed a new IRTree model for answer key generation, which does not include the reason provided for inconclusive responses (as the models in Chapter 3 did). This simplification was made for two reasons: the first is to fit the data onto the ‘conclusiveness’ scale in Figure 4.2 which is used for the CCT model, and the second is that the reasons provided for inconclusive responses are relatively inconsistent. In a follow-up study done by the FBI (Ulery et al., 2012), 72 Black Box study participants were asked to re-assess 25 items. 85% of no value decisions, 90% of exclusion decisions,
68% of inconclusive decisions, and 89% of individualization decisions were repeated; while only 44% of ‘Close’, 21% of ‘Insufficient’, and 51% of ‘No Overlap’ decisions were repeated. Inconclusive reasoning thus varies more within examiners than the source decisions, and a generated answer key containing reasons for inconclusives may not be reliable or consistent across time.

The tree structure for the simplified IRTree is shown in Figure 4.4. The first internal node ($Y^*_{1}$) represents the value assessment, the second internal node ($Y^*_{2}$) represents the conclusive assessment, and the third internal node represents the individualization/exclusion assessment. Note that $Y^*_{3}$ is not a part of the conclusiveness scale in Figure 4.2 and thus provides additional information beyond the ‘conclusiveness’ answer key.

![Figure 4.4: The answer key IRTree](image)

In order to fit an IRTree model, the observed categorical responses, $Y$ are transformed into three binary matrices: $Y^*_{1}$, $Y^*_{2}$, and $Y^*_{3}$. $Y^*_{1}$ represents the matrix for the first split in the tree (where a 1 represents ‘No Value’ and a 0 represents ‘Has Value’), $Y^*_{2}$ represents the matrix for the second split in the tree (1 = Inconclusive, 0 = conclusive), and $Y^*_{3}$ represents the third split (1= individualization, 0 = exclusion).
For example, if the observed responses were represented in matrix form:

\[
Y = \begin{bmatrix}
NA & Individ. & Excl. & \ldots & NV \\
Inconc. & Individ. & NA & \ldots & Excl. \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
NV & NA & Inconc. & \ldots & Excl.
\end{bmatrix},
\]

where the rows correspond to participants and the columns correspond to items, the following three binary matrices would be constructed to represent the data:

\[
Y_1^* = \begin{bmatrix}
NA & 0 & 0 & \ldots & 1 \\
0 & 0 & NA & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & NA & 0 & \ldots & 0
\end{bmatrix},
Y_2^* = \begin{bmatrix}
NA & 0 & 0 & \ldots & NA \\
1 & 0 & NA & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
NA & NA & 1 & \ldots & 0
\end{bmatrix},
Y_3^* = \begin{bmatrix}
NA & 1 & 0 & \ldots & NA \\
NA & NA & \ldots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
NA & NA & NA & \ldots & 0
\end{bmatrix}.
\]

Also define

\[
Z_1 = \begin{cases}
1 & \text{if } Y_1^* \neq NA \\
0 & \text{if } Y_1^* = NA
\end{cases},
Z_2 = \begin{cases}
1 & \text{if } Y_2^* \neq NA \\
0 & \text{if } Y_2^* = NA
\end{cases},
Z_3 = \begin{cases}
1 & \text{if } Y_3^* \neq NA \\
0 & \text{if } Y_3^* = NA
\end{cases}.
\]

\( Y_1^* \), \( Y_2^* \) and \( Y_3^* \) are parameterized using Rasch models, each with distinct latent participant parameters (\( \theta_{i1}, \theta_{i2}, \theta_{i3} \)) and item parameters (\( b_{j1}, b_{j2}, b_{j3} \)). Note that, since the branch splits in the tree do not correspond to correct/incorrect decisions, these latent Rasch parameters do not have the usual proficiency/difficulty interpretations. \( \theta_{ik} \) is instead participant \( i \)'s propensity to choose the left branch at split \( k \). That is, \( \theta_{i1} \) is participant \( i \)'s propensity towards ‘No Value’ decisions, \( \theta_{i2} \) is participant \( i \)'s propensity towards ‘Inconclusive’ decisions, and \( \theta_{i3} \) is participant \( i \)'s propensity towards ‘Individualization’ decisions. The item parameters, \( b_{jk} \), denote the item propensity towards the right branch, so \( b_{j1} \) is item \( j \)'s propensity towards ‘Value’, etc.
The probability of observing each of the possible responses can then be calculated by multiplying the branch probabilities that lead to each leaf. The probability of observing a ‘No value’ response is simply the Rasch model corresponding to the first split:

\[ P(Y_{ij} = \text{No Value}) = P(Y_{ij1}^* = 1) \]
\[ = \text{logit}^{-1}(\theta_{i1} - b_{j1}). \]

(4.18)

(4.19)

While the probability of observing an ‘Inconclusive’ consists of the product of probability of observing a ‘No Value’ (i.e. \(1 - P(Y_{ij} = \text{No Value})\)) and an additional Rasch probability conditional on \(Y_{ij}^* = 0\):

\[ P(Y_{ij} = \text{Inconclusive}) = P(Y_{ij1}^* = 0) \times P(Y_{ij2} = 1|Y_{ij1} = 0) \]
\[ = (1 - \text{logit}^{-1}(\theta_{i1} - b_{j1})) \times \text{logit}^{-1}(\theta_{i2} - b_{j2}). \]

(4.20)

(4.21)

Finally, the probability of an ‘individualization’ and ‘Exclusion’ are given by:

\[ P(Y_{ij} = \text{Individualization}) = P(Y_{ij1}^* = 0) \times P(Y_{ij2} = 0|Y_{ij1} = 0) \times P(Y_{ij3} = 1|Y_{ij2} = 0) \]
\[ = (1 - \text{logit}^{-1}(\theta_{i1} - b_{j1})) \times (1 - \text{logit}^{-1}(\theta_{i2} - b_{j2})) \times \text{logit}^{-1}(\theta_{i3} - b_{j3}). \]

(4.22)

(4.23)

and

\[ P(Y_{ij} = \text{Exclusion}) = P(Y_{ij1}^* = 0) \times P(Y_{ij2} = 0|Y_{ij1} = 0) \times P(Y_{ij3} = 0|Y_{ij2} = 0) \]
\[ = (1 - \text{logit}^{-1}(\theta_{i1} - b_{j1})) \times (1 - \text{logit}^{-1}(\theta_{i2} - b_{j2})) \times (1 - \text{logit}^{-1}(\theta_{i3} - b_{j3})). \]

(4.24)

(4.25)
Assume \( \theta = \{\theta_1, \theta_2, \theta_3\} \sim MVN(0, \Sigma_\theta) \) and \( b = \{b_1, b_2, b_3\} \sim MVN(\mu_b, \Sigma_b) \), which identifies the model by restricting the means of the participant parameters to be zero and does not assume independence between the different dimensions of \( \theta \) and \( b \).

Then, assuming local independence across items and independence between participants, the likelihood is:

\[
L(Y|\theta, b) = \prod_I \prod_J \exp(y_{1ij}z_{1ij}\delta_{1ij}) \exp(y_{2ij}z_{2ij}\delta_{2ij}) \exp(y_{3ij}z_{3ij}\delta_{3ij}) \frac{1 + \exp(\delta_{1ij})}{1 + \exp(\delta_{2ij})}\frac{1 + \exp(\delta_{3ij})}{1 + \exp(\delta_{3ij})}
\]

where \( \delta_{ki} = \theta_{ki} - b_{kj} \).

### 4.4 Sufficient Statistics

The LTRM and IRTree models take different approaches to modeling response behavior, yet obtain similar item-level responses in this setting. This section derives conditional sufficient statistics, or shows that simple (nontrivial) sufficient statistics do not exist, for each model considered. Simple sufficient statistics allow for a better understanding of differences in model estimates and predictions, and provide a theoretical framework for understanding why seemingly unrelated models produce similar results.

#### 4.4.1 Item Response Tree

The likelihood derived in Section 4.3 can be written as:

\[
P(Y_{ij} = y_{ij}|\theta, b) = \prod_I \prod_J \frac{\exp(y_{1ij}z_{1ij}(\theta_{1i} - b_{1j}) + y_{2ij}z_{2ij}(\theta_{2i} - b_{2j}) + y_{3ij}z_{3ij}(\theta_{3i} - b_{3j}))}{[1 + \exp(\theta_{1i} - b_{1j})]^{z_{1ij}}[1 + \exp(\theta_{2i} - b_{2j})]^{z_{2ij}}[1 + \exp(\theta_{3i} - b_{3j})]^{z_{3ij}}}.
\]

Consider a single split in the tree (e.g. \( Y_{3}^* \)). Conditional on other parts of the tree, a Rasch model results. The sufficient statistics for Rasch model parameters are simply the person score or item score.
For example, consider

\[
P(Y_{3i} = y_{3i}|y_1, y_2, z, b, \theta_1, \theta_2) = \prod_j \frac{\exp(y_{i1}z_{1j}(\theta_1 - b_{1j}) + y_{i2}z_{2j}(\theta_2 - b_{2j}) + y_{i3}z_{3j}(\theta_3 - b_{3j}))}{[1 + \exp(\theta_1 - b_{1j})]^{z_{1j}}[1 + \exp(\theta_2 - b_{2j})]^{z_{2j}}[1 + \exp(\theta_3 - b_{3j})]^{z_{3j}}}
\]

\[
= \prod_j \frac{\exp(\sum_j y_{i1}z_{1j}(\theta_1 - b_{1j}) + y_{i2}z_{2j}(\theta_2 - b_{2j}) - y_{i3}z_{3j}b_{3j})}{\prod_j[1 + \exp(\theta_1 - b_{1j})]^{z_{1j}}[1 + \exp(\theta_2 - b_{2j})]^{z_{2j}}[1 + \exp(\theta_3 - b_{3j})]^{z_{3j}}}
\]

\[
= \frac{\exp(\theta_3 \sum_j y_{i3}z_{3j})}{\prod_j[1 + \exp(\theta_3 - b_{3j})]^{z_{3j}} g(Y)}
\]

and so \(\sum_j y_{i3}z_{3j}\) is sufficient for \(\theta_3\), assuming \(y_1, y_2, \theta_1, \theta_2, z, b\) are known. Although it might seem unreasonable to assume \(z_3\) is known when estimating \(\theta_3\), \(z_3\) does not include any additional information after knowing \(y_2\) and \(z_1\). That is, \(z_3 = 1\) if \(z_1 = 1\) or \(y_2 = 1\).

The conditional sufficient statistics for the item parameters, \(b_{kij}\), are therefore the marginal category totals for category \(k\). The sufficient statistic for \(b_{1ij}\) is the number of no value observations for item \(j\), the conditional sufficient statistic for \(b_{2ij}\) is the number of inconclusive observations for item \(j\), and the conditional sufficient statistic for \(b_{3ij}\) is the number of individualizations for item \(j\).

### 4.4.2 LTRM

Suppose \(Y_{ij}\) takes on values \(1, ..., C\). Then:

\[
P(Y_{ij} = y_{ij}) = F_z(a_i \gamma_{c} + b_i) - F_y(a_i \gamma_{c-1} + b_i)
\]

Where \(F_z\) is the CDF of \(Z\), \(\gamma_0 = -\infty\) and \(\gamma_C = \infty\).

As seen in Section 4.3.1 the LTRM is a cumulative probit model. What makes the LTRM different from standard IRT models is that the location parameters \((T_j)\) do not correspond to difficulty, and a separate item ‘difficulty’ parameter \((\lambda_j)\) is included that governs \(\tau_{ij}\) (reporting noise). Additionally,
there are three latent person parameters \((E_i, a_i, b_i)\) instead of a single latent trait. One of these person parameters \((a_i)\) is a scaling term on the category boundaries. Note that the LTRM can be written as:

\[
P(Y_{ij} = y_c) = F(a_i \gamma + b_i) - F(a_i \gamma + b_i) \quad (4.26)
\]

\[
= \Phi\left(\frac{a_i \gamma + b_i - T_k}{\tau_{ik}}\right) - \Phi\left(\frac{a_i \gamma + b_i - T_k}{\tau_{ik}}\right) \quad \text{where } \Phi = \text{CDF of } N(0, 1) \quad (4.27)
\]

\[
= \Phi(b_i - (-\gamma_c) - T_k) - \Phi(b_i - (-\gamma_c - 1) - T_k) \quad \text{assuming } \tau_{ik} = 1, a_i = 1 \quad (4.28)
\]

which has the same form as the Rating Scale version of the Probit Graded Response Model (Samejima, 1969) (i.e. constant item discrimination and constant step parameters across items modeled with a cumulative probit model).

Sijtsma and Hemker (2000) call this a “difference model” and note that neither stochastic ordering or invariant item ordering hold. As noted in Masters (1982), item and person parameters are not separable in this class of models and so simple sufficient statistics for item and person parameters do not exist.

The “latent truth” location of item \(j\) \((T_j)\) is the main parameter of interest for generating “answer keys”. \(T_j\) is also the true mean of the “latent appraisals”, \(Z_{ij}\) (i.e. \(Z_{ij} \sim N(T_j, 1/\tau_{ij})\)). If \(Z_{ij}\) were observed, the sufficient statistic for \(T_j\) would simply be \(\frac{1}{n_j} \sum_{m \in I_j} z_{mj}\), where \(I_j\) is the set of participants who answered question \(j\) and \(n_j = |I_j|\).

Unfortunately, only \(Y_{ij}\) is observed. It is possible, however, to exploit the fact that:

\[
Y_{ij} = \begin{cases} 
1 & Z_{ij} < a_i \gamma_1 + b_i \\
2 & a_i \gamma_1 + b_i \leq Z_{ij} \leq a_i \gamma_2 + b_i \\
3 & Z_{ij} > a_i \gamma_2 + b_i 
\end{cases}
\]

in order to bound \(Z_j\). Start by defining \(\zeta_l\) as a reasonable lower bound for \(Z_{ij}\) and \(\zeta_h\) as a reasonable upper bound for \(Z_{ij}\). Then \(\sum_{i \in I_j} Z_{ij}\) can be bounded above by:
\[
\sum_{i \in I_j} Z_{ij} = \sum_{n_1} Z_{(i)j} + \sum_{n_2} Z_{(i)j} + \sum_{n_3} Z_{(i)j} \tag{4.29}
\]

\[
\leq \sum_{n_1} (a_i \gamma_1 + b_i) Y_{ij} = 1 + \sum_{n_2} (a_i \gamma_2 + b_i) Y_{ij} = 2 + \sum_{n_3} \zeta_h Y_{ij} = 3 \tag{4.30}
\]

\[
\leq \sum_{n_1} (a_{(n)} \gamma_1 + b_{(n)}) Y_{ij} = 1 + \sum_{n_2} (a_{(n)} \gamma_2 + b_{(n)}) Y_{ij} = 2 + \sum_{n_3} \zeta_h Y_{ij} = 3 \tag{4.31}
\]

\[
= (a_{(n)} \gamma_1 + b_{(n)}) \sum_{i \in I_j} 1_{Y_{ij} = 1} + (a_{(n)} \gamma_2 + b_{(n)}) \sum_{i \in I_j} 1_{Y_{ij} = 2} + \zeta_h \sum_{i \in I_j} 1_{Y_{ij} = 3} \tag{4.32}
\]

where \( \sum_{i \in I_j} 1_{Y_{ij} = 1} \) is the number of observations in category 1 for question \( j \), \( \sum_{i \in I_j} 1_{Y_{ij} = 2} \) is the number of observations in category 2 for question \( j \), and \( \sum_{i \in I_j} 1_{Y_{ij} = 3} \) is the number of observations in category 3 for question \( j \).

Similarly, \( \sum_{i \in I_j} Z_{ij} \) can be bounded below by:

\[
\sum_{i \in I_j} Z_{ij} \geq \sum_{n_1} \zeta_i 1_{Y_{ij} = 1} + \sum_{n_2} (a_i \gamma_2 + b_i) 1_{Y_{ij} = 1} + \sum_{n_3} (a_i \gamma_2 + b_i) 1_{Y_{ij} = 3} \tag{4.33}
\]

\[
\geq \sum_{n_1} \zeta_i 1_{Y_{ij} = 1} + \sum_{n_2} (a_{(1)} \gamma_1 + b_{(1)}) 1_{Y_{ij} = 2} + \sum_{n_3} (a_{(1)} \gamma_2 + b_{(1)}) 1_{Y_{ij} = 3} \tag{4.34}
\]

\[
= \zeta \sum_{i \in I_j} 1_{Y_{ij} = 1} + (a_{(1)} \gamma_1 + b_{(1)}) \sum_{i \in I_j} 1_{Y_{ij} = 2} + (a_{(1)} \gamma_2 + b_{(1)}) \sum_{i \in I_j} 1_{Y_{ij} = 3} \tag{4.35}
\]

Therefore, although simple sufficient statistics do not exist for the LTRM, there is at least some relationship between the estimates for \( T_j \) and the marginal totals for each item.

### 4.4.3 Adapted Cumulative-Logits LTRM

Recall that the probability of observing category \( c \) under the Cumulative Categories LTRM is given by:

\[
P(Y_{ij} = c|b_i) = \frac{\exp(b_i - T_j + a_i \gamma_c)}{1 + \exp(b_i - T_j + a_i \gamma_c)} - \frac{\exp(b_i - T_j + a_i \gamma_{c-1})}{1 + \exp(b_i - T_j + a_i \gamma_{c-1})} \tag{4.36}
\]

where \( \gamma_0 = -\infty \) and \( \gamma_c = \infty \).
Consider the probability of observing a response pattern with $J$ items, $P(Y_i = y_i | b_i) = \prod_{j=1}^{J} P(Y_{ij} = y_{ij} | b_i)$, which results in a quantity in which $b_i$ and $T_j$ are not separable. Therefore simple sufficient statistics do not exist.

### 4.4.4 Adapted Adjacent-Logits LTRM

In the adapted adjacent-categories LTRM (A-LTRM), the probability of observing category $c$ is given by:

$$P(Y_{ij} = c | b_i) = \exp(\sum_{k=0}^{c} b_i - T_j + a_i \gamma_k) \over \sum_{l=0}^{m} \exp \sum_{k=0}^{l} b_i - T_j + a_i \gamma_k$$

and so the probability of participant $i$'s response pattern to $J$ items is then:

$$P(Y_i = y_i | b_i) = \prod_{j=1}^{J} P(Y_{ij} = y_{ij} | b_i) = \frac{\exp(\sum_{j=1}^{J} \sum_{k=0}^{c} b_i - T_j + a_i \gamma_k)}{\prod_{j=1}^{J} \sum_{m=0}^{K} \exp \sum_{k=0}^{m} b_i - T_j + a_i \gamma_k}$$

and then

$$\log P(Y_i = y_i | b_i) = b_i \sum_{j=1}^{J} y_{ij} - \sum_{j=1}^{J} y_{ij}T_j + a_i \sum_{k=1}^{Y} \gamma_k - \log(\prod_{j=1}^{J} \sum_{m=0}^{K} \exp \sum_{k=0}^{m} b_i - T_j + a_i \gamma_k).$$

Thus, by the Fisher-Neyman factorization theorem, the observed total score, $\sum_{j=1}^{J} y_{ij}$, is sufficient for $b_i$, conditional on the other parameters. By a similar argument, the item score, $\sum_{i=1}^{I} y_{ij}$, is conditionally sufficient for $T_j$.

### 4.4.5 Summary

There are therefore no simple sufficient statistics, conditional on other parameters, for the LTRM or C-LTRM. The conditional sufficient statistic for the item location parameter, $T_j$, in the A-LTRM is $\sum_{i=1}^{I} y_{ij}$.
the marginal sum for item $j$. The conditional sufficient statistics for the item location parameters in the IRTree model are the marginal category sums, $\sum_{i=1}^{I} y_{kij} z_{kij}$. The sufficient statistics for $T_j$ in the LTRM can be bounded by the marginal category sums, meaning that the IRTree and the LTRM use the same information present in the data to locate the item parameters, even though the models are parameterized in different ways. These results are summarized in Table 4.1.

### Table 4.1: Summary of conditional sufficient statistics for item parameters in the LTRM, A-LTRM, C-LTRM, and IRTree models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sufficient Statistic for Item Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTRM</td>
<td>No simple statistic exists *Bounded by marginal category sums</td>
</tr>
<tr>
<td>C-LTRM</td>
<td>No simple statistic exists</td>
</tr>
<tr>
<td>A-LTRM</td>
<td>$\sum_{i=1}^{I} y_{ij}$ (marginal sum)</td>
</tr>
<tr>
<td>IRTree</td>
<td>$\sum_{i=1}^{I} y_{kij} z_{kij}$ (marginal category sum)</td>
</tr>
</tbody>
</table>

#### 4.5 LTRM and IRTree applied to the FBI “black box” study

Four models were fit to the FBI “black box” dataset: the original LTRM (LTRM), adapted cumulative-categories LTRM (C-LTRM), adapted adjacent-categories LTRM (A-LTRM), and the IRTree model defined above (IRTree).

The LTRM was fit using the R package **CCTpack** (Anders 2017). The **CCTpack** implementation of the LTRM contains a number of informative priors and hyper-priors, many of which differ from the LTRM specified in (Anders and Batchelder 2015) (changes are summarized in Appendix 4.8.1), but a full discussion of the appropriateness of these specifications is omitted here. The A-LTRM and C-LTRM were fit using Stan (Stan Development Team 2018b) with prior distributions specified as in Table 4.2.

The IRTree model was fit using the same priors as the efficient implementation discussed in Chapter 3.
Table 4.2: Comparison of priors in LTRM CCTpack implementation, and the A-LTRM and C-LTRM implementations.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>LTRM Prior (CCTpack)</th>
<th>A-LTRM</th>
<th>C-LTRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_j$</td>
<td>$N(\mu_T, \sigma_T^2)$</td>
<td>$N(\mu_T, \sigma_T^2)$</td>
<td>$N(\mu_T, \sigma_T^2)$</td>
</tr>
<tr>
<td>$\mu_T$</td>
<td>$N(0, 4)$</td>
<td>$N(0, 5)$</td>
<td>$N(0, 5)$</td>
</tr>
<tr>
<td>$\sigma_T^2$</td>
<td>$U(.25, 3)$</td>
<td>Cauchy$(0, 2.5)$</td>
<td>Cauchy$(0, 2.5)$</td>
</tr>
<tr>
<td>$\lambda_j$</td>
<td>$N(\mu_{\lambda}, \sigma_{\lambda}^2)T(-2.3, 2.3)$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\mu_{\lambda}$</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\sigma_{\lambda}^2$</td>
<td>$U(.25, 2)$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\gamma_c$</td>
<td>$N(\mu_\gamma, \sigma_\gamma^2)$</td>
<td>$U(-10, 10)$</td>
<td>$U(-10, 10)$</td>
</tr>
<tr>
<td>$\mu_\gamma$</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\tau_\gamma$</td>
<td>10</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$E_i$</td>
<td>lognormal$(\mu_E, \tau_E)$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\mu_E$</td>
<td>$N(0, 10)$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\tau_E$</td>
<td>$G(.01, .01)$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\alpha_i$</td>
<td>lognormal$(\mu_a, \tau_a)T(-2.3, 2.3)$</td>
<td>lognormal$(\mu_a, \tau_a)$</td>
<td>lognormal$(\mu_a, \tau_a)$</td>
</tr>
<tr>
<td>$\mu_a$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_a$</td>
<td>$G(.01, .01)T(.01,)$</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>$b_i$</td>
<td>$N(\mu_b, \sigma_b^2)$</td>
<td>$N(\mu_b, \sigma_b^2)$</td>
<td>$N(\mu_b, \sigma_b^2)$</td>
</tr>
<tr>
<td>$\mu_b$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma_b^2$</td>
<td>$U(.25, 2)$</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

4.5.1 Model Fit

WAIC (Watanabe 2010) as defined in Vehtari et al. (2017) was used to compare the models to one another. Since the LTRM, A-LTRM, and C-LTRM predict responses as one of no value, inconclusive, or conclusive; while the IRTree distinguishes between individualizations and exclusions, individualizations and exclusions were grouped together in the IRTree model in order to compare all four models directly.

Table 4.3 shows the WAIC, with standard errors as well as in-sample prediction error, for each of the four models. The IRTree model performed the best, the C-LTRM and A-LTRM performed slightly worse but very similarly (within a standard error of one another), and the LTRM performed the worst. Small perturbations in the LTRM parameters (especially $E_i$ and $\lambda_j$, which govern the variance of the latent assessments of item location) can lead to large changes in $P(Y_{ij} = y_{ij})$. Although the LTRM has a fair performance on average (i.e. the in-sample prediction error), looking at the posterior draws individually through WAIC reveals the difficulty in locating parameters in such a large parameter space.
TABLE 4.3: WAIC for each of the four models. In order to compare the IRTree, individualizations and exclusions were grouped together (i.e. $Y^*_j$ was ignored).

<table>
<thead>
<tr>
<th>Model</th>
<th>WAIC</th>
<th>SE</th>
<th>In-Sample Prediction Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTRM</td>
<td>40416</td>
<td>748</td>
<td>0.19</td>
</tr>
<tr>
<td>C-LTRM</td>
<td>13976</td>
<td>175</td>
<td>0.14</td>
</tr>
<tr>
<td>A-LTRM</td>
<td>14053</td>
<td>178</td>
<td>0.15</td>
</tr>
<tr>
<td>IRTree</td>
<td>12484</td>
<td>166</td>
<td>0.12</td>
</tr>
</tbody>
</table>

4.5.2 Item Location estimates compared to Marginal Category Totals

Section 4.4 showed that the sufficient statistics for the IRTree item parameters, conditional on the other model parameters, are the marginal category totals. Additionally, the conditional sufficient statistics for the item parameters in the A-LTRM are a function of the marginal totals. Since the A-LTRM is closely related to both the original LTRM model and the C-LTRM, one would expect a similar relationship between the item parameters and the marginal category totals across models. In this section, the item parameter estimates for each of the four models are compared to the observed number of observations in each category.

Recall that the item parameters of interest are:

- Item location ($T_j$) in the original Latent Truth Rater Model
- Item location ($T_j$) in the Adapted Cumulative-Categories LTRM (C-LTRM)
- Item location ($T_j$) in the Adapted Adjacent-Categories LTRM (A-LTRM)
- Item value tendency ($b_{j1}$) in the IRTree model
- Item conclusive tendency ($b_{j2}$) in the IRTree model
- Item exclusion tendency ($b_{j3}$) in the IRTree model

In order to more easily compare the parameters across models, $T_j$ in the LTRM was multiplied by a scaling factor in order to present all LTRM models on the same scale, and is denoted $T^*$. The IRTree $b_{j1}$ and $b_{j2}$ parameters were also multiplied by $-1$ to correspond to no value and inconclusive propensities, respectively.
Figure 4.5 shows the item location parameters from each of the four models compared to the observed number of no value responses for each item. The LTRM $T^*$, A-LTRM $T$, and C-LTRM $T$ have similar estimates, with more positive estimates being associated with more observed no value responses (as expected). The same relationship is observed in the IRTree $-b_1$ (item no value propensity) and $-b_2$ (item inconclusive propensity) estimates. It makes sense that $b_1$ and $b_2$ would share this relationship, as items that are more likely to be rated as ‘no value’ are also likely to be rated as ‘inconclusive’ if they make it to the next branch in the tree. The $b_3$ estimate in the IRTree corresponds to the individualization/exclusion decision, with higher item parameters being more likely to be rated as an exclusion. The relationship between $b_3$ and the number of no value observations is interesting, as it suggests that items in which more ‘no value’ decisions are observed are more likely to be exclusions than individualizations.

There are also distinct ‘bands’ in the LTRM $T^*$ estimates. Rather than taking many different values within the latent continuum, the item locations are concentrating very close to one of five values: (1) those at the upper truncated limit, which correspond to the items that were unanimously no value, (2) those at the lower truncated limit, which correspond to the items that were unanimously conclusive, (3) a concentration around zero, corresponding to items that were mostly inconclusive, (4) a concentration...
between the upper limit and zero, corresponding to no value items that were not unanimous, (5) a concentration between the lower limit and zero, corresponding to conclusive items that were not unanimous. Similar banding occurs in the other models as well: the top bands in the A-LTRM, C-LTRM, and IRTree $-b_1$ and $-b_2$ correspond to items that were unanimously rated as no value. Appendix 4.8.2 shows the same graphs as Figures 4.5, 4.6, and 4.7 colored by whether responses were unanimous.

![Figure 4.6: Item location parameters from each of the four models compared to the observed number of inconclusive responses for that item. More inconclusive responses are associated with $T$ parameters closer to zero in the LTRM models, higher $b_1$ (value tendency) parameters and lower $b_2$ parameters (conclusive tendency) in the IRTree](image)

Figure 4.6 shows the same parameter estimates but compared to the number of inconclusive responses for each item. All of the LTRM models produce $T_j$ estimates that approach zero as the number of inconclusives increase. Since those models use the latent scale to separate three categories, where inconclusive represents the middle category, it is not surprising that these items are near zero. The banding remains noticeable in the original LTRM model. For the IRTree parameters, more observed inconclusives are generally associated with higher $b_1$ (more likely to be of value) and lower $b_2$ (more likely to be inconclusive).

Figure 4.7 shows the same parameter estimates compared to the number of observed conclusives for each item. More conclusive responses are associated with lower $T$ parameters in the adapted and original LTRM models, as well as lower parameters in IRTree $-b_1$ (value propensity) and $-b_2$ (conclusive propensity). The banding is still present in the original LTRM model, and there is additional grouping.
noticeable in the other estimates as well. For the IRTree $b_3$ estimate, for example, there are three groups of $b_3$ estimates: one for obvious exclusions, one for obvious individualizations, and one for non-obvious conclusive responses. The lower bands in the adapted LTRM models and the IRTree $b_1$ and $b_2$ estimates correspond to the items that were unanimously conclusive (additional figure in Appendix 4.8.2).

Since $b_3$ in the IRTree model contains information about the individualization/exclusion split, which the LTRM models do not, that parameter is also compared to the observed number of individualizations and exclusions. Figure 4.8 shows that $b_1$ and $b_2$ both increase as either individualizations or exclusions increase, corresponding to items that are more likely to be rated as having value and being conclusive. As individualizations increase, $b_3$ decreases and as exclusions increase, $b_3$ increases. Note that there are again two bands in the $b_3$ estimates, corresponding to the items that were unanimous and not unanimous.

### 4.5.3 Generated Answer Key Comparison

One purpose of estimating the item parameters in the LTRM models is to generate an answer key. In this section, the answer keys from the LTRM models are compared to each other and to the answer key from the IRTree model.
Figure 4.8: Item location parameters from each of the four models compared to the observed number of individualization (left) and exclusion (right) responses for that item. As observed individualizations or exclusions increase, \( b_1 \) and \( b_2 \) increase. Higher values of \( b_3 \) are associated with more exclusions, lower values of \( b_3 \) are associated with more individualizations.

As a baseline, the modal response for each question using the observed responses was computed. Unlike the IRTree and LTRM approaches, this answer key does not account for different tendencies of examiners who answered each item; nor does it account for items being answered by different numbers of examiners.

The LTRM, A-LTRM, and C-LTRM all estimate the answer key, a combination of \( T_j \)'s and \( \gamma_c \)'s, directly. The answer for item \( j \) is ‘No Value’ if \( T_j < \gamma_1 \), ‘Inconclusive’ if \( \gamma_1 < T_j < \gamma_2 \) and ‘Conclusive’ if \( T_j > \gamma_2 \).

For the IRTree, an answer key was calculated based on what one would expect an ‘unbiased examiner’ to respond. The response of a hypothetical unbiased examiner (i.e. \( \theta_{ki} = 0 \forall k \)) to each question was predicted, using the estimated item parameters in each split.

Note that for both the LTRM models and the IRTree, one could also justify generating answer keys based on the predictive responses across all examiners. That is, rather than using \( \gamma_1 \) and \( \gamma_2 \) as cutoffs for the LTRM answer keys, the average cutoffs for all examiners could be used: \( \bar{a}\gamma_1 + \bar{b} \) and \( \bar{a}\gamma_2 + \bar{b} \). For the IRTree, one could calculate the modal posterior predictive answer for each question instead of using \( \theta_{ki} = 0 \) to predict an answer. By using \( \gamma_1 \) and \( \gamma_2 \) as cutoffs (i.e. \( a_i = 1 \) and \( b_i = 0 \)) for the LTRM answer keys and \( \theta_{ki} = 0 \) for the IRTree answer key instead, answer keys were generated based on the expected
response for an idealized unbiased examiner. This approach allows for the answer keys for different models to be interpreted in the same way.

There are thus five answer keys:

1. Modal answer key
2. LTRM answer key
3. A-LTRM answer key
4. C-LTRM answer key
5. IRTree answer key

Table 4.4 shows the number of items (out of 744) that the answer keys disagree on. The most similar answer keys are the A-LTRM and C-LTRM, which only disagreed on six items: three that disagreed between inconclusive/conclusive and three that disagreed between no value and inconclusive. The original LTRM model most closely matched the modal answer, with the A-LTRM model disagreeing with the modal answer most often.

<table>
<thead>
<tr>
<th></th>
<th>Modal</th>
<th>LTRM</th>
<th>C-LTRM</th>
<th>A-LTRM</th>
<th>IRTree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modal</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LTRM</td>
<td>12</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C-LTRM</td>
<td>48</td>
<td>39</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A-LTRM</td>
<td>52</td>
<td>43</td>
<td>6</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>IRTree</td>
<td>32</td>
<td>24</td>
<td>28</td>
<td>34</td>
<td>0</td>
</tr>
</tbody>
</table>

There were 48 items for which at least one of the models disagreed with the others. The vast majority of these disagreements were between ‘no value’ and ‘inconclusive’ or ‘inconclusive’ and ‘conclusive’. Of the 48 items in which models disagreed, only five items were rated to be conclusive by some models and no value by others. All of these five items were predicted to be ‘no value’ by the LTRM, ‘inconclusive’ by the A-LTRM and C-LTRM, and ‘exclusion’ by the IRTree. Table 4.5 shows the number
of observed responses in each category for these five items and illuminates two problems with the LTRM approaches. First, the original LTRM strictly follows the modal response, even when a substantial number of participants came to a different conclusion. In Question 665, for example, eight examiners were able to make a correct exclusion, while the LTRM still chose ‘no value’ as the correct response. Second, the A-LTRM and C-LTRM models may rely too much on the ordering of outcomes. Both adapted LTRM models predicted these items to be inconclusives, yet most examiners who saw the items rated it as either a ‘no value’ or ‘exclusion’.

**Table 4.5**: The number of observed responses in each category for the five items with a disagreement between no value and conclusive.

<table>
<thead>
<tr>
<th>Item ID</th>
<th>No Value</th>
<th>Inconclusive</th>
<th>Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>427</td>
<td>13</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>438</td>
<td>12</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>443</td>
<td>7</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>665</td>
<td>9</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>668</td>
<td>14</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>

### 4.5.4 Differences in Person Parameters

In addition to generating answer keys, it is also important to examine how the models assess the participants. In this section, the estimated participant parameters under each of the models are compared. We will first compare the LTRM, A-LTRM, and C-LTRM shift and scale parameters to one another, and then compare the LTRM estimates to the IRTree estimates.

The adapted LTRMs should have similar estimates as the original LTRM for the person shift ($b_i$) and scale ($a_i$) parameters. The original LTRM model, however, also included a person ‘competency’ parameter in the precision term, and so the $b_i$ and $a_i$ parameters have been scaled accordingly. These parameters provide insight into how different examiners use the boundaries between no value/inconclusive/conclusive responses.

Figure 4.9 shows the shift parameters (left) and scale parameters (right) pairs plots for each of the LTRM models. As expected, there are only small differences in estimated person parameters between
the C-LTRM and A-LTRM. The adjusted shift parameter \((b_i \times E_i)\) in the original LTRM is much smaller in magnitude than the adapted LTRM shift parameters (-0.05 to 0.5 as compared to -4 to 4). The adjusted scale parameter \((a_i \times E_i)\) in the original LTRM also appears to be completely unrelated to the scale parameter in the adapted models.

The relative noisiness of the original LTRM parameters compared to the adapted LTRM parameters could be due to the item difficulty parameter \((\lambda_j)\) being inseparable from the person parameters. (Recall that \(\tau_{ij} = E_i / \lambda_j\), where \(\tau_{ij}\) is the precision of the latent appraisals of \(T_j\).) It is also possible that the LTRM is over-parameterized in this case, and there is not enough information present in the data to locate each parameter.

![Figure 4.9: Person shift \(b_i\) and scale \(a_i\) parameters from the original LTRM (LTRM), cumulative-categories LTRM (C-LTRM), and adjacent-categories LTRM (A-LTRM). The C-LTRM and A-LTRM estimates are virtually the same and much larger in scale than the original LTRM.](image)

The IRTree model takes an entirely different approach, modeling propensity towards each decision (no value, inconclusive, exclusion) as a distinct process. Some relationship between \(b_1\) and \(b_2\) should thus be expected in the IRTree and the shift and scale parameters in the LTRMs.

Figure 4.10 shows the person shift and scale parameters from the A-LTRM and original LTRM compared to the IRTree estimates for each person parameter (Inconclusive propensity, individualization propensity, and no value propensity). The original LTRM shift and scale parameters were again transformed to account for \(E_i\). The same comparison to the C-LTRM is omitted since the estimated parameters are essentially the same as the A-LTRM.
There is a weak relationship between the shift parameters from the LTRM and the parameters from the IRTree model (bottom row). The shift and scale parameters from the A-LTRM have a stronger relationship with the IRTree parameters. In particular, examiners that are more likely to report inconclusives tend to have lower shift parameters and larger scale parameters, while examiners that are more likely to report no values tend to have lower shift parameters and smaller scale parameters. There is likely more information about examiner tendencies encoded in the shift and scale parameters in the adapted LTRM than in the original LTRM.

![Figure 4.10: Person shift $b_i$ (left) and scale $a_i$ (right) parameters from the original LTRM (LTRM) and adjacent-categories LTRM (A-LTRM) compared to the IRTree person parameters. There is a stronger relationship between the A-LTRM and IRTree parameters than the original LTRM and IRTree parameters.](image)

As mentioned above, the adapted LTRM models did not include a person ‘competency’ parameter (denoted $E_i$ in the original LTRM). It is possible that examiner tendencies are encoded in $E_i$ instead of $a_i$ or $b_i$ in the LTRM model. Figure 4.11 shows $E_i$ plotted against the IRTree parameters. The $E_i$’s are very small in magnitude (.20 to .35), and do not appear to be related to the IRTree parameters.

The IRTree person parameters have the benefit of being readily interpretable, with justification for the estimates clearly visible in the data through the conditional sufficient statistics (marginal person category totals). Based on a comparison to the IRTree parameters, the LTRM does not have any advantage over the adapted LTRMs, which do not include a person competency or item difficulty parameter.
4.5.5 Logs comparison of C-LTRM and A-LTRM

Finally, the category probability curves for the A-LTRM and C-LTRM are compared to one another. Figure 4.12 shows the category probability curves for eight evenly-spaced items. For the same person on the same item, the cumulative categories model will predict slightly more inconclusive responses and slightly fewer no value responses than the adjacent categories model. However, as noted in Section 4.5.3 the A-LTRM and C-LTRM disagree on only six questions, which are equally split between disagreeing on conclusive/inconclusive and disagreeing on inconclusive/no value. This difference therefore has little practical effect on model predictions.

Figure 4.12: Comparison of category probabilities for eight evenly-spaced question locations for the A-LTRM (ac) and C-LTRM (cc). The probability curves for conclusive are the same, while there is a slight separation of the curves for the inconclusive and no value categories.
4.6 Discussion

The LTRM provides an interesting framework to conceptualize responses to items for which an answer key is unknown. When applied to this setting, however, where item location on the latent scale is closely related to item difficulty, it is likely over-parameterized. There is not enough information in the data for the various person parameters (i.e. shift, scale, competency) to be meaningful. We proposed two new models that removed the participant competency and item difficulty parameters from the LTRM: an adapted LTRM modeled with cumulative-logits (C-LTRM) and adapted LTRM modeled with adjacent logits (A-LTRM). The adapted models led to useful, identified parameter estimates. By including a ‘scale’ parameter for each person, the category thresholds are allowed to vary by person, a novel way to assess reporting differences. Using an adjacent-categories logit (A-LTRM) model instead of a cumulative categories logit (C-LTRM) model leads to simple conditional sufficient statistics for the item location and person reporting shift parameters. Although there are small differences in the logit curves under each model, these differences do not lead to many practical changes in model inferences or predictions. However, both of the adapted models may over-rely on the ordering of categories.

The IRTree model, on the other hand, does not require the responses to be ordered, which allows for explicit modeling of both ‘no value’ decisions and exclusion/individualization decisions. Incorporating ‘no value’, ‘individualization’, and ‘exclusion’ decisions would require a new scale under the LTRM framework. The person and item parameters under the IRTree model are readily interpretable and are clearly explainable from the observed data, and simple conditional sufficient statistics exist for these parameters. Furthermore, a model-based approach such as the one developed in Chapter 3 could be used to determine whether inconclusive responses are justified or mistaken.

Using a model-based framework to generate expected answers provides more robust answer keys than relying on the observed responses alone. Both IRTrees and a CCT-based approach allow for the estimation of person and item effects alongside an answer key. Furthermore, although the two approaches are formulated quite differently, they lead to similar generated answer keys in the Black...
Box data. This similarity is due to the conditional sufficient statistics for item location parameters being closely related in the two models. For this setting, we prefer using the IRTree framework to analyze responses because it does not require the responses to be ordered and because each decision may be modeled explicitly.

Generating evidence to be used for assessment purposes is both time-consuming and difficult. The methods introduced in this chapter provide a way to use evidence collected in non-controlled settings, for which ground truth is unknown, for assessment purposes.

In practice, answer keys should be generated for both proficiency tests and error-rate studies. Examiners should receive feedback not only when they make false identifications or exclusions, but also when mistaken ‘no value’ or ‘inconclusive’ decisions are made. It is therefore important to distinguish when no value, inconclusive, individualization, and exclusion responses are expected in a forensic analysis. As seen in the CTS data earlier in this dissertation (Section 2.3), however, participants should not be automatically punished for failing to match consensus and/or expected reasons. Rather, the generated answer keys should be used to flag items and/or participants for further study.

4.7 Conclusions

How can response data for which there is an unknown correct answer be modeled within the IRT framework?

This chapter developed an IRTree framework to use test responses, without additional information about the items, to generate an expected ‘answer key’ that does not depend on examiner tendencies. This allows response data for which there is a correct answer, but the correct answer is unknown, to be modeled within the IRTree framework. The LTRM, an existing approach to ‘IRT without an answer key’ for ordinal responses, was compared to an answer key generated from an IRTree model. The approaches produced similar answer keys and were shown to have closely-related conditional sufficient statistics.
For this setting, we prefer using the IRTree framework to analyze responses because it does not require
the responses to be ordered and because each decision may be modeled explicitly.

**When participants are allowed to specify an uncertain response, how can it be determined if that
uncertainty is justified or a mistake?**

The generated answer keys from either the LTRM or the IRTree specify when inconclusive responses are
expected. One approach to identifying mistaken uncertain responses is to simply ‘score’ the data based
on the generated answer key. A second approach was introduced in Chapter 3 in which the estimated
probabilities of each response under the model can also be calculated, and uncertain responses marked
as mistaken if the probability of a conclusive response is sufficiently high.
### 4.8 Appendix

#### 4.8.1 CCTpack implementation of LTRM

TABLE 4.6: Comparison of priors in LTRM specifications using JAGS conventions for parameterization (e.g. the normal distribution is parameterized with precision instead of variance). $T(l, h)$ denotes truncating a random variable to be between $l$ and $h$. Changes in distribution between the paper (Anders and Batchelder, 2015) and the software package (Anders et al., 2017) are boldfaced.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior (Paper)</th>
<th>Prior (CCTpack)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_j$</td>
<td>$N(\mu_T, \tau_T)$</td>
<td>$N(\mu_T, \tau_T)$</td>
</tr>
<tr>
<td>$\mu_T$</td>
<td>$N(0, .1)$</td>
<td>$N(0, .25)$</td>
</tr>
<tr>
<td>$\sigma_T$</td>
<td>-</td>
<td>$U(.25, 3)$</td>
</tr>
<tr>
<td>$\tau_T$</td>
<td>$G(1, .1)$</td>
<td>-</td>
</tr>
<tr>
<td>$\lambda_j$</td>
<td>$G(\mu_E \lambda^2 \tau_\lambda, \mu_\lambda, \tau_\lambda)$</td>
<td>$N(\mu_\lambda, \tau_\lambda) T(-2.3, 2.3)$</td>
</tr>
<tr>
<td>$\mu_\lambda$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma_\lambda$</td>
<td>-</td>
<td>$U(.25, 2)$</td>
</tr>
<tr>
<td>$\tau_\lambda$</td>
<td>$G(1, .1)$</td>
<td>-</td>
</tr>
<tr>
<td>$\gamma_c$</td>
<td>$N(\mu_\gamma, \tau_\gamma)$</td>
<td>$N(\mu_\gamma, \tau_\gamma)$</td>
</tr>
<tr>
<td>$\mu_\gamma$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_\gamma$</td>
<td>0.1</td>
<td>.1</td>
</tr>
<tr>
<td>$E_i$</td>
<td>$G(\mu_E^2 \tau_E, \mu_E \tau_E)$</td>
<td>lognormal$(\mu_E, \tau_E)$</td>
</tr>
<tr>
<td>$\mu_E$</td>
<td>$G(4, 4)$</td>
<td>$N(0, .1)$</td>
</tr>
<tr>
<td>$\tau_E$</td>
<td>$G(4, 4)$</td>
<td>$G(.01, .01)$</td>
</tr>
<tr>
<td>$a_i$</td>
<td>$G(\mu_a^2 \tau_a, \mu_a \tau_a)$</td>
<td>lognormal$(\mu_a, \tau_a) T(-2.3, 2.3)$</td>
</tr>
<tr>
<td>$\mu_a$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_a$</td>
<td>$G(1, .1)$</td>
<td>$G(.01, .01) T(.01, .01)$</td>
</tr>
<tr>
<td>$b_i$</td>
<td>$N(\mu_b, \tau_b)$</td>
<td>$N(\mu_b, \tau_b)$</td>
</tr>
<tr>
<td>$\mu_b$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_b$</td>
<td>$G(1, .1)$</td>
<td>$U(.25, 2)$</td>
</tr>
</tbody>
</table>
4.8.2 Banding in the item parameters

Figure 4.13: Item parameters for each of the four models against three observed quantities: the number of conclusives ($n_{conc}$, top row), the number of inconclusives ($n_{inc}$, middle row), and the number of no values ($n_{nv}$, bottom row). Unanimous items are shown in color based on which response was observed for that item.
Chapter 5

Incorporating Self-Reported Collateral Responses

In many testing situations, including forensic decision-making studies and proficiency tests, it is increasingly common to collect collateral information alongside responses to each item. This collateral information may include the time spent on each question, participant confidence in their answer, or perceived item difficulty. There are therefore multiple responses for each participant \( \times \) item pair that might depend on shared latent variables. For instance, reported difficulty may be related to both participant proficiency and ‘true’ item difficulty. This chapter focuses on the following questions:

1. Can modeling additional self-reports (e.g. reported difficulty) lead to a better understanding of the implications of making casework decisions based on such reports?

2. Can reported difficulty be incorporated directly into a model for responses?

3. Can responses be used to model differential use of the subjective reporting scales?
5.1 Introduction

In addition to the value assessment and source conclusion for each item, which have been the focus of previous chapters of this dissertation, the FBI black box study also asked participants to report the difficulty of each item they evaluated on a five-point scale. These reported difficulties are not the purpose of the assessment, but are related to the responses and collected at the same time and can therefore be thought of as ‘collateral information’. In forensic science settings, other collateral information could include: covariates describing the evidence (e.g. quality, number of features, etc.), covariates describing examiner background (e.g. years experience, type of training, etc.), examiner confidence, or a subjective likelihood ratio. In this chapter, a framework is proposed for modeling collateral information in general.

There is a rich literature on a number of different approaches for incorporating variables that are not the response variable into an IRT analysis. When the additional variables are covariates describing either the participants or the items, they can be used as predictors for proficiency or difficulty in the IRT model (de Boeck and Wilson 2004). For instance, image quality or number of minutiae may be predictive of item difficulty in fingerprint analysis, and experience and type of training may explain examiner proficiency. A common example of an explanatory IRT model is the Linear Logistic Test Model (LLTM), which uses item properties to model difficulty with a linear model. A further approach is Differential Item Functioning (DIF), which represents the additional covariate as an interaction between an item indicator and a person predictor representing group membership (Holland and Wainer 2012). Covariates describing the participants could also be used for Latent Class Analysis (LCA), which divides the population of participants into a finite number of mutually exclusive classes (Hagenaars and McCutcheon 2002).

Additional variables can also be incorporated when they represent responses, rather than covariates. Nylund-Gibson et al. (2019) first characterize latent classes, with the latent class then predicting further outcomes. Whittaker et al. (2012) provide an analysis of model selection for comparing tests with multiple types of item responses (e.g. multiple choice and open-ended). Repeated responses from the same participant to the same item are also used for longitudinal studies. For instance, Barbieri et al.
model multiple responses over time to measure impact of treatment on health-related quality of life.

There are two reasons why these common approaches are not appropriate for incorporating reported difficulty into an IRT analysis of the Black Box Study. The first reason is that reported difficulty is not a covariate describing either items or participants. Rather it is a second subjective evaluation of the item, and thus depends on both participant and item characteristics. The second reason is that the source decision and reported difficulty measure fundamentally different aspects of the task. The source evaluation can be represented as a correct or incorrect decision, but there is no clear way to represent reported difficulty as a correct/incorrect decision.

This chapter takes the approach of using a single latent variable model for multiple observed variable types. Although reported difficulty is not often measured alongside responses in educational testing or other standard IRT settings, response time is often used as a proxy for difficulty, under the assumption that larger response times correspond to higher perceived difficulty. Thissen (1983) provides an early example of this type of modeling, where the logarithm of response time is modeled as a linear function of the standard IRT quantity (proficiency - difficulty) and additional latent variables for both items and participants. Ferrando and Lorenzo-Seva (2007), van der Linden (2006) each propose various models for modeling response time jointly with the traditional correct/incorrect IRT response. Modeling collateral information alongside responses in such a way has been shown to improve estimates of IRT parameters through the sharing of information (van der Linden et al., 2010).

Examiners likely use different thresholds for coming to source decisions (AAAS, 2017; Ulery et al., 2017), and may use different standards for reporting difficulty as well. These differences could also be attributed to differences in typical casework, experience, or use of the reporting scale. Reported difficulty also provides additional information about the items beyond standard IRT estimates. For example, consider two items with identical response patterns (i.e. the same examiners answered each question correctly and incorrectly) but one item was reported to be more difficult than the other by all examiners. It is completely plausible that most examiners struggled with one of those items, but eventually came
to the correct conclusion. Standard IRT will not detect the relative difficulty of that item compared to
the truly easier item with the same response pattern. Incorporating reported difficulty into a model
explicitly could help identify such items.

While the framework proposed in this chapter can be used to draw further inferences about
examiners and/or items, it could also be used for other types of collateral responses. This framework
could be especially useful for cases in which examiners are asked to report a confidence or subjective
likelihood ratio regarding a source decision. If evidence is discarded or further analysis is not pursued
based on examiner evaluation of their confidence or the difficulty of the evidence, differences in
reporting among examiners could be especially consequential.

This chapter models reported difficulty alongside IRT for value and source decisions. Section 5.3
introduces a joint modeling approach using IRT, with an extension for the IRTree setting introduced in
Section 5.4. Section 5.5 discusses results of the modeling approach applied to the FBI black box data, and
Section 5.6 summarizes conclusions.

5.2 Reported Difficulty in the FBI Black Box Study

Recall from Chapter 2 that each examiner × item response in the FBI Black Box Data contains:

• Mating: whether the pair of prints were “Mates” (a match) or “Non-mates” (a non-match)

• 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 (one of ‘Close’, ‘Insuffi-
  cient’, or ‘No Overlap’)

• Exclusion Reason: If exclusion, the reason for the exclusion (one of ‘Minutiae’ or ‘Pattern’)

Therefore, in addition to participant × item evaluation decisions, which are some function of Mating, Latent\_Value, Compare\_Value, and Inconclusive\_Value depending on the scoring scheme and denoted $Y_{ij}$ throughout this thesis, the FBI black box study also recorded reported difficulty (denoted $X_{ij}$ in this thesis) on a five point scale. That is, for each examiner × item response in the data, examiner $i$ rated item $j$ as one of: A-Obvious, B-Easy, C-Medium, D-Difficult, E-Very Difficult. When examiners reported a ‘no value’ decision for $Y_{ij}$, no response is recorded for reported difficulty.

### 5.3 Joint Model of Responses and Reported Difficulty for Scored Data

#### 5.3.1 Empirical Motivation

In both lenient and harsh scoring schemes, there is a relationship between reported difficulty and correctness in the Black Box study. Consider a simple logistic regression model predicting correctness, $Y_n$, where $n$ indexes an item × examiner response, using reported difficulty alone:

$$P(Y_n = 1) = \logit^{-1}(D_n\beta)$$

where $D_n = \{1_{X_n=A}, 1_{X_n=B}, 1_{X_n=C}, 1_{X_n=D}, 1_{X_n=E}\}$, and $X_n$ is the reported difficulty for observation $n$ (ranges from A=Obvious to E=Very Difficult). Under three different scoring schemes from Chapter 2, the predicted $P$(Correct) is noticeably different across at least two reported difficulty based on 95% confidence intervals (Figure 5.1a).

Additionally, reported difficulty is related to the “correctness propensity”, $\delta_{ij} = \theta_i - b_j$, from a standard Rasch model under different scoring schemes (Figure 5.1b). Estimated IRT quantities (participant proficiency and item difficulty) are therefore related to reported difficulty as well as correctness, and could perhaps be used to model both responses. Modeling reported difficulty alongside responses also
allows for the assessment of reporting differences across participants. Similarly, incorporating reported difficulty could identify items that are consistently reported to be more or less difficult than their responses suggest.

(A) Predicted \( P(\text{Correct}) \) (with 95% confidence intervals) under three scoring methods, based only on reported difficulty.

(B) Rasch correctness propensity \( \theta_i - b_j \) for each examiner \( \times \) item pair under each scoring scheme by reported difficulty.

Figure 5.1: Empirical evidence that reported difficulty is related to both correctness (A) and Rasch correctness propensity (B)

### 5.3.2 Joint Model for Standard IRT

Let \( Y_{ij} \) be the scored response of participant \( i \) to item \( j \), and let \( X_{ij} \) be the difficulty reported by participant \( i \) to item \( j \). \( Y_{ij} \) thus takes the values 0 (incorrect) or 1 (correct), and \( X_{ij} \) is an ordered categorical variable with five levels (A=obvious to E=very difficult).

One option for incorporating \( X_{ij} \) is to include it as a participant \( \times \) item covariate in the IRT model (e.g. \( \logit(P(Y_{ij} = 1)) = \theta_i - b_j + D_{ij}\beta \) (as above)). While variations on such a model are possible, this chapter takes an approach that models reported difficulty as a distinct response. Treating reported difficulty as an additional response allows for differences across examiners and/or items in reporting tendencies. Additionally, reported difficulty is collected immediately following the evaluation of the item, while covariates (or predictors) are generally known before test administration. It therefore makes sense to treat reported difficulty as a separate response, since it does not exist before the item is evaluated.

We combine a Rasch model:
logit\(P(Y_{ij} = 1) = \theta_i - b_j\) 

(5.1)

with a cumulative-logits ordered logistic model for the reported difficulties:

\[X^*_{ij} = \logit^{-1} f(\theta_i, b_j, \zeta),\]

(5.2)

where \(\zeta\) are possible additional latent variables, and

\[
X_{ij} = \begin{cases} 
\text{A-Obvious} & X^*_{ij} \leq \gamma_1 \\
\text{B-Easy} & \gamma_1 < X^*_{ij} \leq \gamma_2 \\
\text{C-Medium} & \gamma_2 < X^*_{ij} \leq \gamma_3 \\
\text{D-Difficult} & \gamma_3 < X^*_{ij} \leq \gamma_4 \\
\text{E-Very Difficult} & X^*_{ij} > \gamma_4 
\end{cases}
\]  

(5.3)

As discussed in Chapter 4, there are alternative approaches for modeling ordered categorical responses. We chose a cumulative-logits approach because it is directly implemented in Stan and therefore runs slightly faster than an adjacent-logits or other approaches. In Chapter 4 there was no practical effect observed on modeling outcomes between an adjacent-logits and cumulative-logits approach. We have no reason to believe this choice has a practical effect on modeling outcomes here either. Other formulations could certainly be used.

We consider the following models for \(X^*_{ij}\) (corresponding to different functions \(f\) and additional latent variables \(\zeta\)):

1. Common Slope and Intercept for all participants and items (common SI): \(f(\theta_i, b_j, \zeta) = g \cdot (\theta_i - b_j) + h\)

2. Common Intercept, individual slopes for each participant (common intercept): \(f(\theta_i, b_j, \zeta) = g_i \cdot (\theta_i - b_j) + h\)
3. Common Slope, individual intercepts for each participant (common slope): \( f(\theta_i, b_j, \zeta) = g \cdot (\theta_i - b_j) + h_i \)

4. Individual slopes and intercepts for each participant (free SI): \( f(\theta_i, b_j, \zeta) = g_i \cdot (\theta_i - b_j) + h_i \)

5. No intercept, slope for each item (item slope): \( f(\theta_i, b_j, \zeta) = g_j \cdot (\theta_i - b_j) \)

6. Common slope, individual intercept for each participant and each item (Thissen (1983) model adaptation) (thissen): \( f(\theta_i, b_j, \zeta) = g \cdot (\theta_i - b_j) + h_i + f_j \)

Additional latent variables allow for the assessment of whether examiners are biased in how they use the scale. For example, examiners with \( h_i > 0 \) tend to over-report difficulty relative to expectations (based on examiner proficiency \( \theta_i \)), estimated item difficulty \( b_j \), and other examiners’ reported difficulty for item \( j \) \( (f_j) \), while examiners with \( h_i < 0 \) tend to under-report difficulty relative to expectations. Additionally, it can determined if there are items whose difficulty is not captured in other estimated effects (e.g. \( f_i > 0 \) or \( f_i < 0 \)).

We assume that (1) each participant’s responses are independent of other participants’ responses \( (Y_i \perp Y_i') \), (2) within-participant responses and reports are conditionally independent of one another given the latent trait \( Y_{ij} \perp Y_{ij}'|\theta_i \) and \( X_{ij} \perp X_{ij}'|\theta_i, h_i \), and (3) responses are conditionally independent of reported difficulty given all latent variables \( (X_{ij} \perp Y_{ij}|\theta_i, b_j, \zeta) \). Then the likelihood is:

\[
L(Y, X|\theta, b, \zeta) = \prod_i \prod_j P(Y_{ij} = 1)^{Y_{ij}} (1 - P(Y_{ij} = 1))^{1-Y_{ij}} P(X_{ij} = x_{ij}) \tag{5.4}
\]

and

\[
P(X_{ij} = c) = P(\text{logit}^{-1} f(\theta_i, b_j, \zeta) \leq \gamma_c) - P(\text{logit}^{-1} f(\theta_i, b_j, \zeta) \leq \gamma_{c-1}), \tag{5.5}
\]

where \( \gamma_0 = -\infty \) and \( \gamma_5 = \infty \).
5.4 Joint Model for IRTree Responses and Reported Difficulty

The models presented above are only appropriate for binary $Y_{ij}$ responses that represent correct and incorrect decisions. As an alternative, the raw responses (e.g. \textit{LatentValue} and \textit{CompareValue} from the FBI Black Box Study) could be modeled with an IRTree, and the estimated branch propensities could be used to predict reported difficulty. As shown in Section 5.5.2, using the same parameters for multiple responses can lead to difficulties in estimation, particularly when there are few observations in a given response category. For this reason, the simplified version of the IRTree from Chapter 4 was used, which includes the value decision and source decision, but not inconclusive reasons (reproduced in Figure 5.2 below).

![Figure 5.2: The item response tree](image)

Let $\delta_{1ij} = \theta_{1i} - b_{1j}$, $\delta_{2ij} = \theta_{2i} - b_{2j}$, and $\delta_{3ij} = \theta_{3i} - b_{3j}$, where $\theta_{ki}$ corresponds to the $i^{th}$ person parameter at split $k$ and $b_{kj}$ corresponds to the $j^{th}$ item parameter at split $k$. The IRTree probability model is then:

\begin{align*}
P(Y_{ij} = \text{No Value}) &= \logit^{-1}(\delta_{1ij}) \\
P(Y_{ij} = \text{Inconclusive}) &= (1 - \logit^{-1}(\delta_{1ij}))\logit^{-1}(\delta_{2ij}) \\
P(Y_{ij} = \text{Individualization}) &= (1 - \logit^{-1}(\delta_{1ij}))(1 - \logit^{-1}(\delta_{2ij}))\logit^{-1}(\delta_{3ij}) \\
P(Y_{ij} = \text{Exclusion}) &= (1 - \logit^{-1}(\delta_{1ij}))(1 - \logit^{-1}(\delta_{2ij}))(1 - \logit^{-1}(\delta_{3ij})).
\end{align*}
5.4.1 Empirical Motivation

Section 5.1 demonstrated that reported difficulty was related to correctness propensity \( \theta_i - b_j \) estimated from a Rasch model. In the IRTree model, since \( \theta \) and \( b \) no longer correspond to proficiency and difficulty, but to person and item tendencies towards different decisions, the same relationships between \( \theta_i, b_j \) and \( X_{ij} \) (reported difficulty) are unlikely to hold.

Instead, reported difficulty \( (X_{ij}) \) tends to increase as both No Value propensity and Inconclusive propensity increase (Figure 5.3). That is, items that are more likely to be determined to be inconclusive or not of value are generally reported to be more difficult. Additionally, items with individualization propensity \( (\theta_3 - b_3) \) closer to zero tend to be reported as more difficult than items with extreme individualization propensity.

![Figure 5.3: Estimated IRTree decision propensities \( (\theta_{ki} - b_{kj}) \) plotted by reported difficulty (A (Obvious) to E (Very Difficult)) for each Examiner \( \times \) Item response in the data. 'NA' reported difficulties correspond to 'No Value' responses. Reported difficulty has the strongest relationship with the no value and inconclusive decision propensities.](image)

A natural next question is whether any of the estimated IRTree parameters stochastically order reported difficulty. First, it was investigated whether the mean reported difficulty can be ordered by any of the IRTree parameters. If the mean is not ordered by a parameter, then the observations can not be ordered by that parameter.
The participant parameters \( (\theta) \) and item parameters \( (b) \) do not order the reported difficulties (See Appendix), but \( \delta_{kij}'s \) (recall \( \delta_{kij} = \theta_{ki} - b_{kj} \)) do order the means of reported difficulties. Since \( \delta_{kij} \) does not correspond to correctness propensity in the IRTree case, but to a tendency towards one decision or another, it would make sense that small \( \delta_k 's \) (in magnitude) would be more “difficult” (i.e. no clear tendency towards either split) than large (in magnitude) \( \delta_{kij}'s \). For this reason, \( |\delta_{kij}|'s \) are also examined.

Figure 5.4 shows the average reported difficulty compared to the average \( \delta_{kij} \) (or \( |\delta_{kij}| \)) for fifty groups binned by \( \delta_{kij} \) or \( |\delta_{kij}| \) (where points are colored by group). That is, all observations are ordered by either \( \delta_{kij} \) or \( |\delta_{kij}| \) and then divided into fifty equally-sized groups. Within each group, the average reported difficulty and average \( \delta_{kij} \) (or \( |\delta_{kij}| \)) is calculated. If \( \delta_{kij} \) or \( |\delta_{kij}| \) stochastically orders reported difficulty, there should be a strictly increasing or decreasing trend.

![Figure 5.4](image)

**Figure 5.4:** Average reported difficulty (within-bin) against average \( \delta_{kij} \) (or \( |\delta_{kij}| \)) for fifty groups of \((i,j)\) observations binned by \( \delta_{kij} \) (or \( |\delta_{kij}| \)). Groups with all missing observations for reported difficulty are plotted with \( \bar{x} = 0 \) (four groups in first panel).

Although average reported difficulty appears to have some relationship with all of the \( \delta_k 's \) and \( |\delta_k|'s \), the relationship is most pronounced in \( |\delta_2| \). That is, observations with \( |\delta_2| \) near zero (no clear tendency towards inconclusive or conclusive) tend to be reported as most difficult, while those with a very clear tendency towards inconclusive or conclusive tend to be reported as easiest.
The cumulative frequency (analogous to empirical CDF) of reported difficulties for each binned group is shown in Figure 5.5, where the colors of the groups correspond to the same group colors as Figure 5.4. Since the ‘No Value’ observations don’t have a reported difficulty, and therefore do not clearly fit in the reported difficulty ordering from A (obvious) to E (very difficult), treating those observations in one of three ways is explored: (1) Remove them (first plot), (2) treat them as the highest possible difficulty (second plot), (3) treat them as the lowest possible difficulty (third plot). Options (1) and (3) lead to nearly stochastically ordered distributions.

![Figure 5.5: Cumulative frequency of reported difficulties for each $|\delta^2|$ bin from Figure 5.4, bottom middle panel.](image)

Figures 5.4 and 5.5 suggest there is a relationship between the estimated IRTree propensities, particularly $\delta_{2ij} = \theta_{2i} - b_{2j}$, and reported difficulty. The estimated IRTree quantities could thus likely be used in a joint model for responses and reported difficulty.

### 5.4.2 Joint Model for IRTree

We combine an IRTree Model (Equation 5.9) with a cumulative-logits ordered logistic regression model for the rated difficulties, as in Section 5.3. Five models were fit to test whether $\delta_{kij}$, or $|\delta_{kij}|$, predict reported difficulty sufficiently (models i and ii), or if additional latent parameters are needed (models
A model using $|\delta_{2ij}|$ as the only predictor (model v) is also included, based on the strong relationship observed in Figures 5.4 and 5.5.

i. IRTree: $X_{ij}^* = \logit^{-1} \sum_{k=1}^{3} \beta_k \delta_{kij}$

ii. Absolute value IRTree: $X_{ij}^* = \logit^{-1} \sum_{k=1}^{3} \beta_k |\delta_{kij}|$

iii. Thissen-IRTree: $X_{ij}^* = \logit^{-1} \sum_{k=1}^{3} \beta_k \delta_{kij} + h_i + f_j$

iv. Thissen-Absolute value IRTree: $X_{ij}^* = \logit^{-1} \sum_{k=1}^{3} \beta_k |\delta_{kij}| + h_i + f_j$

v. Absolute $\delta_2$ IRTree: $X_{ij}^* = \logit^{-1} \beta_2 |\delta_{2ij}|$

where

$$X_{ij} = \begin{cases} 
A-Obvious & X_{ij}^* \leq \gamma_1 \\
B-Easy & \gamma_1 < X_{ij}^* \leq \gamma_2 \\
C-Medium & \gamma_2 < X_{ij}^* \leq \gamma_3 \\
D-Difficult & \gamma_3 < X_{ij}^* \leq \gamma_4 \\
E-Very Difficult & X_{ij}^* > \gamma_4 
\end{cases}$$

### 5.5 Results

#### 5.5.1 Joint Model for Standard IRT

As discussed throughout this dissertation, results from applying Rasch models to the FBI black box study are largely dependent on the scoring scheme used. The six models discussed in Section 5.3 were therefore fit to the black box data under two scoring schemes: treating the inconclusives as missing completely at random (Inconclusive MCAR) and scored under a consensus-based scoring scheme (Consensus).
As in earlier chapters, the widely-applicable information criterion (WAIC, Watanabe (2010)) as defined in (Vehtari et al., 2017) is used to compare the model fit for the \((Y_{ij}, X_{ij})\) responses within each scoring scheme:

\[
\text{WAIC} = \sum_{n=1}^{N} \log \left( \frac{1}{S} \sum_{s=1}^{S} p(y_{n}, x_{n} | \theta^{*}, b^{*}, \zeta^{e}) \right) - 2 \times \left( \hat{l}_{pd} - \sum_{n=1}^{N} V_{s=1}^{S} \log (p(y_{n}, x_{n} | \theta^{*}, b^{*}, \zeta^{e})) \right) \tag{5.10}
\]

where \(V_{s=1}^{S}\) represents the sample variance, \(\theta^{*}, b^{*}, \zeta^{e}\) are the parameter estimates from the \(s^{th}\) posterior draw, and \(\hat{l}_{pd} = \sum_{n=1}^{N} \log \left( \frac{1}{S} \sum_{s=1}^{S} p(y_{n}, x_{n} | \theta^{*}, b^{*}, \zeta^{e}) \right)\) is the log pointwise predictive density.

Table 5.1 shows the WAIC and standard errors for each of the six models fit to the two scoring schemes. For both the Inconclusive MCAR scored data and the Consensus-based scored data, the Thissen model outperforms all of the other models for reported difficulty. Perhaps unsurprisingly, including additional participant and item effects for reported difficulty leads to the best model fit. Additionally, WAIC is substantially lower for the Inconclusive MCAR scoring scheme than for the Consensus-based scoring scheme because the inconclusive MCAR scheme treats much of the data as missing, so there are fewer data points contributing to the WAIC calculation.

In addition to WAIC, each model was also evaluated using a posterior predictive check. For each participant in the dataset, their observed score under the relevant scoring scheme, \(\frac{1}{n_{i}} \sum_{j \in J_{i}} y_{ij}\), and their predicted score under the model, \(\frac{1}{n_{i}} \sum_{j \in J_{i}} \hat{y}_{ij}\), were calculated. If the model is performing well, the predicted scores should be very similar to the observed scores. In general, predicted score is not very informative about IRT model fit, since even poorly-fitting models will predict the observed score well as
long as item difficulties are related to proportion correct (Sinha et al., 2006). In this case, however, IRT parameters were used to predict reported difficulty as well as correctness. Models without additional latent variables to predict reported difficulty may lead to the item parameters being far from the truth and therefore no longer predicting score well.

The predicted scores compared to the observed scores for both scoring schemes are shown in Figure 5.6. As expected, models that do not include additional person-level additive parameters for reported difficulty (common-intercept, common-si, item-slope) no longer predict observed score well. The Thissen model and the common slope model best predict participant scores, and a better predictive performance is observed for the consensus scoring scheme than for the inconclusive MCAR scoring scheme.

Similarly, the performance of the models in predicting reported difficulties was also evaluated with a posterior predictive check. Instead of predicted and observed score, however, predicted and observed average reported difficulty were used, where the observed average reported difficulty is $\frac{1}{n_i} \sum_{j \in I_i} x_{ij}$ and the predicted average reported difficulty is $\frac{1}{n_i} \sum_{j \in I_i} \hat{x}_{ij}$. The best performance is again
observed in the Thissen and common-slope models, but now a slightly better performance is observed in the Inconclusive MCAR scored data than in the consensus-scored data. Reported difficulties for inconclusive responses are also treated as MCAR under this scoring scheme. These reported difficulties are likely more variable across examiners or items than conclusive responses and thus harder to accurately predict, leading to an inferior model fit.

For both scoring schemes, the model that includes additional terms for both items and participants (Thissen-inspired) performs the best, in terms of both WAIC and posterior predictive checks for the person scores. Figure 5.8 shows the proficiency estimates from the Thissen model against the Rasch proficiency estimates (i.e. the model for correctness from Chapter 2 without modeling reported difficulty) for each scoring scheme. The Thissen proficiency estimates do not differ substantially from the Rasch proficiency estimates, although there is a slight shrinkage towards zero of the Thissen proficiency estimates for the inconclusive MCAR scored data. Figure 5.9 shows the item difficulty estimates from the Thissen model against the item difficulty estimates from the Rasch model. Like proficiency estimates, the Thissen difficulties do not differ substantially from the Rasch difficulties. This is due to the inclusion...
of the $h_i$ and $f_j$ parameters for the reported difficulty part of the model, which sufficiently explains the variation in reported difficulty without impacting the IRT parameters.

Recall that the Thissen model predicts reported difficulty as $g \cdot (\theta_i - b_j) + h_i + f_j$. In addition to proficiency and difficulty, “reporting bias” parameters for participants ($h_i$) and items ($f_j$) are also included. Positive $h_i$ and $f_j$ thus increase the expected reported difficulty while negative $h_i$ and $f_j$ decrease the expected reported difficulty.

$h_i$ can thus be interpreted as participant $i$’s tendency to over or under-report difficulty, after accounting for their proficiency ($\theta_i$), difficulty of items they saw ($b_j$’s), and how other participants rated the difficulty of those items ($f_j$’s). Figure 5.10 shows the $h_i$ estimates and 95% posterior intervals for both the consensus-based and inconclusive MCAR scored data compared to proficiency estimates. Positive estimates correspond to examiners that tend to over-report difficulty, while negative estimates correspond to examiners that tend to under-report difficulty. Since there are plenty of observations whose 95% posterior intervals do not overlap with zero, Figure 5.10 provides evidence that there exist differences among examiners in the way they report difficulty. This reporting bias does not appear to have any relationship with the model-based proficiency estimates. That is, examiners who report items to be more difficult (positive $h_i$) do not perform worse than examiners who report items to be easier (negative $h_i$).

Similarly, $f_j$ can be interpreted as item $j$’s tendency to be over or under-reported, after accounting for difficulty, participant proficiency, and participant reporting bias. Figure 5.11 shows the $f_j$ estimates and
95% posterior intervals for both the consensus-based and inconclusive MCAR scored data compared to the difficulty estimates. Positive estimates correspond to items that tend to have over-reported difficulty, while negative estimates correspond to items that tend to have under-reported difficulty. There are a substantial number of items whose posterior intervals do not overlap with zero, providing evidence that there are items which are consistently reported as more or less difficult than responses suggest. Additionally, there is a mild arc-shaped relationship between $f_j$ and $b_j$: items with estimated difficulties near zero are most likely to have over-reported difficulty, and items with very negative or very positive estimated difficulties (corresponding to items that examiners did very poorly or very well on, respectively) tend to have under-reported difficulty.

Incorporating additional latent parameters for reporting biases leads to the best model fit, and does not result in proficiency or difficulty parameters that substantially differ from a Rasch model alone. Additionally, the new latent parameters allow for the identification of participants that over or under report-difficulty, and items that tend to have their difficulty over or under-reported.
Figure 5.11: Item reporting bias \( (f_j) \) with 95% posterior intervals from the Thissen model for each of the scoring schemes compared to item difficulty \( (b_j) \). Items with large posterior intervals correspond to items that were unanimously no value or inconclusive.

5.5.2 Joint Model for IRTree

Recall that the IRTree in Equation 5.9 above is combined with an ordered logistic model for reported difficulty:

i. IRTree: \( X^*_{ij} = \text{logit}^{-1} \sum_{k=1}^{3} \beta_k \delta_{kij} \)

ii. Absolute value IRTree: \( X^*_{ij} = \text{logit}^{-1} \sum_{k=1}^{3} |\beta_k| \delta_{kij} \)

iii. Thissen-IRTree: \( X^*_{ij} = \text{logit}^{-1} \sum_{k=1}^{3} \beta_k \delta_{kij} + h_i + f_j \)

iv. Thissen-Absolute value IRTree: \( X^*_{ij} = \text{logit}^{-1} \sum_{k=1}^{3} |\beta_k| \delta_{kij} + h_i + f_j \)

v. Absolute \( \delta_2 \) IRTree: \( X^*_{ij} = \text{logit}^{-1} |\beta_2| \delta_{2ij} \)

where

\[
X_{ij} = \begin{cases} 
A-Obvious & X^*_{ij} \leq \gamma_1 \\
B-Easy & \gamma_1 < X^*_{ij} \leq \gamma_2 \\
C-Medium & \gamma_2 < X^*_{ij} \leq \gamma_3 \\
D-Difficult & \gamma_3 < X^*_{ij} \leq \gamma_4 \\
E-Very Difficult & X^*_{ij} > \gamma_4 
\end{cases}
\]
The IRTree and Absolute Value IRTree models predict reported difficulty using the same parameters as those that predict responses \( \delta_{kij} = \theta_{ki} - b_{kj} \) along with a scaling term \( \beta_k \). The Thissen-IRTree and Thissen-Absolute value IRTree, on the other hand, use both the IRTree parameters \( \delta_{kij} \) and additional item and person parameters for reported difficulty \( f_j \) and \( h_i \), respectively. The differences between the IRTree models and the Thissen models thus represent a trade off: models with additional item and person parameters likely fit the data better, but models without the additional parameters may use the information already present in the IRTree responses more efficiently.

Based on the empirical analysis in Section 5.4, \( \delta_1 \) and \( \delta_2 \) were expected to be most predictive of reported difficulty, and therefore \( \beta_1 \) and \( \beta_2 \) were expected to be nonzero, with \( \beta_3 \) expected to be near zero. Instead, some identifiability problems were observed. This can be seen in the posterior distributions for \( \beta_k \) in each model in shown in Figure 5.12. The IRTree and Absolute Value IRTree models result in multi-modal posterior distributions for \( \beta_k \) with both positive and negative modes. The model therefore can’t locate a \( \beta_k \) that provides a best fit.

Surprisingly, \( \beta_3 \) is sometimes estimated to be nonzero under the IRTree and Absolute value IRTree models, even though there was no empirical evidence that estimated \( \delta_{3ij} \) was predictive of reported difficulty. This is due to the structure of the IRTree not being accounted for in the models for reported difficulty. Consider an item, \( i' \), that was unanimously rated to be inconclusive. If \( Y_{ij} \) is modeled with an IRTree, there will be no information present in the data to locate \( b_{3j'} \) (since \( b_{3j'} \) corresponds to the exclusion/individualization split for item \( j' \)) and \( b_{3j'} \) won’t be related to reported difficulty and therefore \( \beta_3 \) will be zero. In the model that predicts \( X_{ij} \) alongside \( Y_{ij} \), on the other hand, the information in \( X_{ij'} \) can be used to locate \( b_{3j'} \), leading to an informative \( b_{3j'} \) and therefore a nonzero \( \beta_3 \). In the IRTree and Absolute Value IRTree models, \( b_{3j} \) is likely taking on some of the variation that is included in \( f_j \) in the Thissen models.

Since the IRTree and Absolute Value IRTree are not able to locate unique estimates of \( \beta_k \), we exclude those models from further analysis.
The Thissen and Absolute Value Thissen models, which do include \( f \) and \( j \) estimates for reported difficulty, no longer detect any relationship between the IRTree parameters and reported difficulty. This is evidenced in Figure 5.12 by the Thissen and Absolute Value Thissen posteriors for \( \beta_k \) being very concentrated at zero. Therefore, although identified \( \beta_k \) parameters are obtained, the strong relationship between \( \delta_{2ij} \) and reported difficulty that was observed in Section 5.4 is “washed out” by the additional parameters.

The final model that was fit was the Absolute \( \delta_2 \) IRTree, which only uses \(|\delta_2|\) as a predictor for reported difficulty (i.e. Model 5). Figure 5.12 shows a nonzero, concentrated negative \( \beta_2 \) posterior for this model. In other words, the more ‘sure’ the model is on the \( Y^*_2 \) split (inconclusive vs conclusive), the easier the question is reported to be by participants.

The \( \theta_2 \) estimates, and 95% posterior intervals, for the \(|\delta_2|\)-only IRTree model, Absolute Value Thissen IRTree, and Thissen IRTree are shown Figure 5.13 compared to the \( \theta_2 \) estimates from the IRTree model without including reported difficulty. The \(|\delta_2|\)-only IRTree model and Absolute Value Thissen IRTree result in slightly different estimates of \( \theta_2 \), as evidenced by some intervals being off the diagonal. The 95% posterior interval length for the original IRTree are included as thin lines, and Figure 5.13 also shows that incorporating reported difficulty in the \(|\delta_2|\)-only IRTree model and Absolute Value Thissen IRTree leads to a more precise estimate of \( \theta_2 \).
We also compare the WAIC for the three models with stable posteriors - the Thissen Model, Absolute Value Thissen Model and $|\delta_2|$ Model (Table 5.2). The Absolute Value Thissen IRTree best predicted the responses, followed by the Thissen IRTree, followed by the $|\delta_2|$ only model.

**TABLE 5.2: WAIC for the three models for which unimodal posteriors are obtained.**

<table>
<thead>
<tr>
<th>Model</th>
<th>WAIC</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abs Val Thissen</td>
<td>40776</td>
<td>304</td>
</tr>
<tr>
<td>Thissen</td>
<td>41810</td>
<td>301</td>
</tr>
<tr>
<td>$</td>
<td>\delta_2</td>
<td>$ only</td>
</tr>
</tbody>
</table>

Finally, the $h$ and $f$ estimates in the two Thissen-inspired models are compared to the corresponding $\theta$ and $b$ estimates, respectively. There are no clear relationships between $h$ and $\theta_k$ for any of $k = 1, 2, 3$ (Figure 5.14), nor are there any clear relationships between $f$ and $b_k$ (Figure 5.15).
In forensic science, collateral information is often collected alongside responses for each examiner × item pair. In the FBI Black Box Study, reported difficulty was collected as collateral information, but other examples could include confidence or a subjective likelihood ratio. I have proposed a modeling approach for collateral responses using a single latent variable model for multiple observed variables.

As repeatedly raised as an issue in this thesis, any method which requires data to be scored as correct or incorrect leads to conclusions that are dependent on that scoring scheme. For this reason, an extension for modeling reported difficulty which makes use of the IRTree framework was also provided. An IRTree
approach to modeling responses allows for the assessment of individual differences among examiners and items at multiple steps in the decision-making process, rather than treating the response as a single decision.

Although reported difficulty appears to be related to estimated IRTree propensities, incorporating additional latent variables led to a better fit than using the IRTree parameters alone. Additionally, estimating the joint IRTree model without additional latent variables specific to reported difficulty leads to multi-modal posterior distributions. In models with unimodal posterior estimates, items were likely to be rated as more difficult when the IRTree model was unsure whether to predict the item as inconclusive or conclusive.

Each model that estimated additional participant ($h_i$) and item ($f_j$) parameters to predict reported difficulty found many of those estimates to be nonzero. Examiners with $h_i > 0$ are likely to over-report difficulty after controlling for proficiency, difficulty, and other participants’ reported difficulty, while examiners with $h_i < 0$ are likely to under-report difficulty. Similar effects were found across items, suggesting that there may be properties of the items which make examiners perceive the question to be more difficult (or easy) than performance suggests. If decisions that could impact casework (e.g., discarding evidence or further investigation) are made based on such collateral information, this variation needs to be taken into account.

A drawback of the modeling approach taken here is that reported difficulty was assumed to be independent of the responses, conditional on the latent variables. This conditional independence assumption could be relaxed in future work. Additionally, a more principled approach to model selection for latent variables for a second response variable could be developed.

In practice, collecting collateral information (reported difficulty or otherwise) could provide additional insight into the differences in perception and decision-making than responses alone. For instance, two items with identical response patterns (and thus identical IRT item difficulties) with different reported difficulties could identify features that are commonly thought to make comparisons more difficult, but don’t actually have an effect on performance. Additional training could therefore improve
efficiency by helping examiners identify those features in casework. Similarly, two items with very different response patterns (and therefore very different IRT item difficulties) with identical reported difficulties could identify new features that impact performance that examiners may not be aware of.

Modeling collateral responses explicitly also allows for the assessment of differences in reporting across examiners. While differences in reported difficulty across examiners are interesting, differences in other types of collateral information that may directly impact investigations or prosecutions, such as confidence or a subjective likelihood ratio, would be especially consequential. As a further example, if evidence is discarded based on examiner confidence or reported difficulty, differences in reporting could substantially impact case outcomes.

5.7 Conclusions

How can additional self-reported information be used to better understand the implications of making casework decisions based on such reports?

The models used in this chapter demonstrated substantial variability in reporting tendencies between examiners. Some examiners were found likely to over-report difficulty after controlling for proficiency, difficulty, and other participants’ reported difficulty; others were likely to under-report. If decisions that could impact casework are made based on collateral information, this variation needs to be taken into account. Further collateral responses should be collected in both controlled experimental settings and in casework to best understand the reporting differences.

Can reported difficulty be incorporated directly into a model for responses?

I have proposed a modeling approach for collateral responses using a single latent variable model for multiple observed variables. An extension for modeling reported difficulty which makes use of the IRTTree framework was also provided, which does not require the data to be scored.
Can responses be used to model differential use of the subjective reporting scales?

While this chapter has demonstrated substantial variability in reported difficulty across both examiners and items, differential use of subjective reporting scales was not addressed explicitly and is left as a future area of work. One approach is to let the category thresholds in Equation 5.3 vary by participant, much like the models in Chapter 4. A second approach would be to incorporate reported difficulty as a covariate in the IRT model with a random participant effect.
5.8 Appendix

\[ \theta_1 - b_1 \text{ (No Value)} \quad \theta_2 - b_2 \text{ (Inconc)} \quad \theta_3 - b_3 \text{ (Individ)} \]

\[ b_1 \quad b_2 \quad b_3 \]

\[ \theta_1 \quad \theta_2 \quad \theta_3 \]

**Figure 5.16:** IRTree estimates for each Examiner \( \times \) Item response in the data by reported difficulty
FIGURE 5.17: Average reported difficulty against item parameters

FIGURE 5.18: Average reported difficulty against item parameters
Chapter 6

Summary, Future Directions, and Broader Implications

This dissertation has adapted and implemented statistical models for complex decision-making in the forensic science domain. Chapter 2 provided an overview of the role that human decision-making plays in forensic science, with a focus on fingerprint analysis. In addition to providing examples of how IRT can be used in existing proficiency tests, through a summary of Luby and Kadane (2018), and in error rate studies, through a summary of Luby (2019), Chapter 2 also identified areas where adaptations to standard IRT are needed. Each of the remaining chapters addressed one such issue, using the Black Box study as motivation.

6.1 Dissertation Summary

The first issue addressed was forensic decision-making data often being recorded as a series of sequential responses. Most Item Response Models were designed for binary scored responses or other standard response schemes consisting of a single measurement. In forensic tasks in general, and the Black Box study in particular, examiners were allowed to report an ‘inconclusive’ or ‘no value’
determination. As an alternative to scoring inconclusive or no value responses, each decision could be left intact and the decision process could be analyzed as a whole. Chapter 3 introduced the IRTrees framework (De Boeck and Partchev, 2012) as a method to model sequential decisions in the FBI “black box” study. After comparing multiple IRTree models, we found that a hypothetical binary decision tree led to both the best fit and most interpretable parameters. A model-based method to assess whether inconclusive responses were erroneous or not was also developed.

Chapter 4 expands upon the idea of identifying erroneous inconclusive responses developed in Chapter 3 by using an IRTree model to generate an answer key based on expected answers from an unbiased examiner for each item. We focused on generating an answer key for four outcomes: no value, inconclusive, exclusion, and individualization; which together with the information provided by the FBI provide a complete answer key. The resulting answer key was similar to the answer key generated from existing methods for ‘IRT without an answer key’, while providing a better fit to the data.

In addition to the latent evaluation and source decision, examiners also reported a difficulty after each item. These reported difficulties varied substantially, even when examiners largely agreed on the latent value and source decision. Chapter 5 addresses this variability by modeling reported difficulty as a collateral response. Results suggest that there are some examiners who over or under-report difficulty after adjusting for proficiency, item difficulty, and how other examiners reported the difficulty of the item. Further study is needed to determine how these differences in reporting may impact casework. Similar effects were found in the items, with some items reported to be more or less difficult than the responses alone suggest. Additional examination of these items could identify features that affect reporting of difficulty but do not have an impact on performance.

6.2 Contributions

The first main contribution of this dissertation is the application of Item Response Theory to a new domain: forensic decision-making tasks. Although differences in individual decision-making, and in
the difficulty of items, has been studied in the forensics literature, this dissertation provides a more complete framework for analyzing complex forensic evaluation decisions. While we have focused on decision-making in fingerprint analysis, using the Black Box study as an example, proficiency testing is used in all forensic fields and IRT could be directly applicable in each of those settings. There is a rich literature of statistical methods for analyzing responses with Item Response Theory. Framing forensic decision-making as an item-response analysis makes well-developed statistical methods, theory, and tools available.

Second, this dissertation provides a methodological contribution to the IRT literature. Item Response Trees are relatively new, and I have developed a Bayesian implementation of a wide variety of these models. These models have included both binary and polytomous response models, item covariates, and multiple responses. Examples of model evaluation and comparison through WAIC and posterior predictive checks were also provided. Furthermore, Item Response Trees have not yet been used explicitly to generate answer keys, and this dissertation has provided an example as well as a theoretical link between Item Response Trees and Cultural Consensus Theory, an existing approach to generating answer keys.

6.3 Future Work

While this thesis has covered methods for working with sequential responses (Chapter 3), unknown ground truth (Chapter 4), and self-reported collateral responses (Chapter 5), there are additional issues raised in this thesis that have not been completed. Future work in this line of research is outlined below.

6.3.1 Additional analyses on related data

The FBI recently (June 2019) made additional datasets to myself and other forensic statistics researchers that are related to the Black Box Study (Buscaglia 2019). While some summary results from the new data were included in this thesis as justification for modeling decisions, immediate future work involves
further incorporating this new data into existing work as well as performing new analyses. There are three newly-available datasets: (1) Image quality metrics for the Black Box study, (2) results from a Black Box Repeatability study, and (3) results from a “White Box” study.

**Black Box Study Item Covariates**

The first newly-released dataset consists of automatically calculated quality and clarity values for each of the latent prints used in the Black Box study. These quantities were computed using the LQMetric software, which is a part of the FBI’s *Universal Latent Workstation* software version 6.5 and later. The included quantities are:

- **Overall Quality**: Estimated probability of an AFIS hit
- **Est VID**: Estimated probability that an examiner would assess the latent of value for identification
- **Est VCMP**: Estimated probability that an examiner would assess the latent of value for comparison
- **Overall Clarity**: An image quality measure, calculated as described in [Hicklin et al.](2013).

A natural first step is to incorporate these quantities into the IRT and IRTrees models as additional item covariates. For example, we could model item effects as:

\[
b_i = \beta_0 + \beta_{OQ} OQ_i + \beta_{VID} VID_i + \beta_{VCMP} VCMP_i + \beta_{OC} OC_i + \epsilon, \tag{6.1}
\]

where \( OQ \) = Overall Quality, \( VID \) = Est VID, \( VCMP \) = Est VCMP, and \( OC \) = Overall Clarity. Implementing this approach in the IRTrees framework could identify which automatically calculated quantities are most predictive of item tendencies at each step in the tree.

Also note that two of the quantities (Est VID and Est VCMP) are directly related to the ‘no value’ split in the IRTrees used in Chapters 3, 4, and 5. In particular, recall the estimated \( b_1 \) is the item propensity...
to be rated as ‘of value’ (more positive $b_1$ implies more likely to be of value, more negative $b_1$ implies more likely to be of no value), which is measuring the same quantity as $\text{Est V CMP}$. Figure 6.1 shows $b_1$ from the IRTree approach, $\text{Est V CMP}$, and $\text{Est V ID}$ compared to the actual percentage of ‘has value’ decisions observed for each item, with $b_1$ far and away providing the best measure of ‘has value’ item tendency. This suggests that (a) predicting examiner responses from images alone is quite hard (which is what $\text{Est V CMP}$, and $\text{Est V ID}$ attempt to do), and (b) $\text{Est V CMP}$, and $\text{Est V ID}$ could perhaps be improved through the incorporation of an IRT or IRTree model into the calculation. A further discussion of incorporating images in an IRT model is provided in Section 6.3.2.

![Figure 6.1: $b_1$ from the IRTree approach, $\text{Est V CMP}$, and $\text{Est V ID}$ compared to the actual percentage of ‘has value’ decisions observed for each item. $b_1$ is more closely related to the observed percentage of ‘has value’ decisions, which is the quantity all three estimates are trying to predict.](image)

**Black Box Repeatability Study**

The Black Box Repeatability study (Ulery et al., 2012) recruited 72 of the original Black Box Study participants to re-analyze 25 image pairs that they had seen in the Black Box Study. The dataset thus contains a second response for $72 \times 25 = 1800$ examiner $\times$ item observations that appeared in the Black Box Study.

We have used the repeatability data as justification for using a simpler IRTree model in Chapters 3 and 5, but this data also provides a unique opportunity for model validation. Obtaining out-of-sample prediction error generally requires the use of a hold-out set or an approximation. Throughout this thesis,
we have used WAIC as a measure of out-of-sample prediction error. A second approach to validation, which is impossible in most contexts, would be to use the model estimates from fitting the model to the Black Box data to predict responses on the repeatability study.

This data could also inform the work in Chapter 5 through providing a measure of how stable reported confidence is over time, and perhaps incorporating this additional variation into a formal measurement model.

**White Box Study**

The Black Box study was designed to measure examiner performance, without attempting to determine how those decisions were made (i.e. treating examiner decisions as a “black box” classifier). The White Box study, on the other hand, asked examiners to annotate features, image clarity, and correspondences between latent and reference images when making their determinations.

Published results have included measuring how much information is needed in order to come to an individualization conclusion (Ulery et al., 2014), variations in features identified across examiners (Ulery et al., 2016), how marked features change between the analysis and comparison stage (Ulery et al., 2015), and factors associated with exclusion decisions (Ulery et al., 2017). As was the case in the Black Box study, participants responded to different sets of prints, and taking an IRT or IRTree approach would allow for the separation of examiner and item effects.

The additional information in the image quality assessments and minutiae markup could be incorporated into an IRTree model. Much like reported difficulty (e.g. Chapter 5) these subjective measures likely vary based on both items and participants. Incorporating them into an IRTree model could inform (a) evaluation decisions, (b) reported difficulty, and (c) thresholds for decisions.

We could also use cluster analysis or latent class analysis to analyze the White Box study to determine if there is evidence of distinct groups among examiners or items. If there are groups among examiners or items that appear to use a decision process similar to each other but different than those out of the group,
we could incorporate multiple IRTrees into an analysis. These results could determine the topology of the IRTree and if a mixture-of-IRTrees would better explain the responses.

6.3.2 Images as Covariates

Observed variables that describe the participants or the items in an assessment are also often incorporated into Item Response Models (de Boeck and Wilson, 2004). The Linear Logistic Test Model (LLTM) incorporates properties of the items into a Rasch model to explain differences in the items’ estimated difficulties. That is,

\[ P(Y_{ij} = 1) = \logit^{-1}(\theta_i - \beta X_j), \]

where \( X \) is an \( i \times (k+1) \) matrix in which the columns represent covariates for the item properties and \( \beta \) are linear regression weights. In cases where there are many items, but few item covariates, the LLTM substantially reduces the number of parameters that must be estimated. The Random Effects LLTM includes both a structural component, which is determined by a linear combination of item covariates (as in the LLTM) and an item-specific deviation. That is, \( b_j \sim N(\beta X_j, \sigma^2_b) \).

Oftentimes in forensic science, as in the Black Box study, the items are actually images (or pairs of images). Although determining which properties of the image(s) make evidence more difficult or easier to analyze could substantially improve our understanding of examiner decision-making, it is difficult to translate images into meaningful covariates. In order to use an LLTM approach, informative covariates could be generated from the images in one of two ways. Either latent print experts could hand-label fingerprints with covariates deemed to be relevant, or predictive models could be built to automatically label the images. Hand-labeling fingerprints is time-consuming, and there is no guarantee that chosen covariates would adequately explain performance. On the other hand, Figure 6.1 in Section 6.3.1 demonstrated that automatically-generated covariates may not necessarily generalize to new data.
Alternatively, image analysis techniques could be directly incorporated into the IRT model. That is, rather than $X_i$ representing informative covariates such as an image quality score, number of minutiae, etc., $X_i$ could represent a low-dimensional signature of image $i$. Standard off-the-shelf principal components analysis (PCA), matrix factorization (MF), or linear discriminant analysis (LDA) may prove useful for this approach.

Due to IRB constraints, images from the FBI black box data are not available. The newly-released latent quality metrics or White Box study results may provide useful insight into how informative covariates impact the difficulty of questions, but an approach using the images directly is not possible.

Collaborative Testing Services (CTS), introduced in Chapter 2, has provided copies of twelve latent print comparison proficiency exams administered between 2012 and 2017. Since this data contains both responses and images, both informative covariate generation and image analysis techniques could be explored. Unfortunately, the data is still limited by the number of items (12 exams of 12 questions each = 144 images of unknown prints) and the response format (value assessments and inconclusive decisions were not collected). Furthermore, these tests are designed as competency exams and so there is very little variation in the responses, and the items represent a very small range of potential difficulties. There are very few items with more than a few incorrect responses, and any features in the images that we might find to predict difficulty could just be noise.

Data should be collected that is both information-rich in the responses (as in the Black Box study) and in the items (like the CTS data) in order for these proposed approaches to be most useful.

6.3.3 Applications in Eyewitness Identification

Individual differences among decision-makers and tasks also exist in eyewitness identification and lineups. Although investigative teams have control over lineup design, there are many external factors that play a role in eyewitness identifications, a large number of which can be attributed to differences among eyewitnesses. These differences can be difficult to measure and impossible to control for in many situations.
Eyewitness identification research has largely focused on system variables (i.e. those under the control of the administrator). These variables include instructions that bias witnesses toward a certain decision (Clark, 2005, 2012; Steblay, 1997; Steblay et al., 2014), whether lineups are presented simultaneously (all possible choices are shown at once) or sequentially (choices are shown one at a time) (Clark, 2012; Steblay et al., 2001, 2011), and how visually similar the fillers (known innocent people) included in the lineup are to the suspect (Clark, 2012; Clark and Godfrey, 2009; Fitzgerald et al., 2013).

Variables that cannot be controlled by the administrator, known as estimator variables in the literature, have also been explored in laboratory studies. These include how long the witness was able to see the perpetrator during the crime (Bornstein et al., 2012), how much time has elapsed between the crime and the lineup (Deffenbacher et al., 2008), and race and age biases (Meissner and Brigham, 2001; Rhodes and Anastasi, 2012).

It has been shown that individual differences, such as memory or face-recognition abilities, affect performance in lineup tasks (Andersen et al., 2014), but experiments largely focus on studying system variables in the lineup setting. While these studies are undoubtedly valuable, the effects of system variables after accounting for individual differences among eyewitnesses is perhaps more important. Luby et al. (2018) has proposed the use of Item Response Theory to include both individual differences and system variables (as item covariates) when analyzing experimental lineup data.

The Innocence Project and the National Research Council recommend that lineup administrators ask witnesses to report their confidence after making a lineup decision. There may also be individual differences in how eyewitnesses are using the confidence scale. If some eyewitnesses use a more conservative threshold of reporting than others, the observed confidence levels may not be a stable method of assessing accuracy of a given eyewitness decision. A model similar to those proposed in Chapter 5, could provide insight into how eyewitnesses report confidence.
6.3.4 Applications in Statistics Education Research

IRT has a long history of being used to analyze concept inventories, and statistics is no exception (Libarkin and Anderson, 2005; Zieffler et al., 2010; Wang and Bao, 2010). As statistics education moves towards a more data-centric approach, with a focus on data cleaning, visualization, and analysis, assessments are becoming more open-ended. This shift, while pedagogically sound, also presents a challenge in analyzing assessment results.

In particular, open-ended data analysis tasks are increasingly common. In these types of assessments, students are provided with a dataset and tasked with answering questions about that data through either a guided analysis task or through an analysis of their own choosing. While an instructor might have an idea of what an appropriate analysis looks like, there are a generally a number of pre-processing and modeling decisions that could be considered “correct”. Instead of attempting to score these tasks, an IRTrees approach could be used to maintain the sequential nature of decisions and consider answers conditional on preceding decisions.

Collateral information (e.g. response time, confidence) is also increasingly collected alongside assessment responses. This provides a further application for the modeling approach introduced in Chapter 5 and also provides an opportunity for new approaches. As one example, the research group at Carnegie Mellon has collected response time and participant confidence alongside responses to a concept inventory (In Prep.) with the goal of using the reported confidence to identify misconceptions. Selecting a wrong answer may not signal a misconception if students admit to guessing, while students who select the wrong answer with high confidence may have a misconception.

6.4 Conclusion

In a recent report, the President’s Council of Advisors on Science and Technology recommended increased “black box” error rate studies for subjective forensic methods (PCAST 2016). While this thesis has focused on statistical methods for analyzing latent print examination, and the FBI Black Box study
in particular, each issue we addressed is broadly applicable in subjective forensic science domains. To conclude this dissertation, I summarize my main findings with respect to (a) the FBI Black Box study, (b) latent print examination in general, and (c) the forensic science domain.

### 6.4.1 Results from the Black Box Study

The FBI Black Box study was the first large-scale study designed to measure the accuracy and reliability of latent print examiners’ decision. Although there was a very low false positive rate (0.1%) and mild false negative rate (7.5%) overall, error rates for individual examiners were not provided and are not meaningful since each examiner saw a different set of items.

Item Response Theory, which accounts for both participant proficiency and item difficulty, provides a better picture of performance than error rates alone. IRTrees provide even more useful insights into how examiners and items vary, and which steps in the decision-making process are most affected by examiner differences. For instance, in Chapter 3, the decision between ‘no value’ and ‘has value’ and the decision between ‘individualization’ and ‘close’ was found to have the most variation among examiners. The IRTree person parameters were also found to be related to traditional proficiency estimated from scored data, and therefore responses need not be scored in order to determine the relative performance of examiners.

Although there are plenty of items with substantial variation in the responses, most items were found to have a clear expected answer (under multiple models) in Chapter 4. Examiners should receive feedback not only when they make false identifications or exclusions, but also when mistaken ‘no value’ or ‘inconclusive’ decisions are made. In order to provide such information, expected answers must first be generated.

Reported difficulty was also collected as collateral information in the Black Box study. In Chapter 5, we found that there is evidence that some examiners over-reported difficulty and some examiners under-reported difficulty. There were also some items that were consistently rated as more or less difficult than performance on those items suggest. Reported difficulty is most closely related to
the decision between inconclusive and conclusive decisions: if a (latent, reference) pair is clearly
inconclusive or clearly conclusive, examiners generally rated it as less difficult than (latent, reference)
pairs that were not clearly inconclusive or conclusive.

6.4.2 Implications for Latent Print Examination

The current standards for decision-making in latent print examination are known as ACE-V: Analysis,
Comparison, Evaluation, and Verification. The Black Box study assessed only one of these steps directly
(the analysis step corresponds to the latent print evaluation decision). The comparison and evaluation
step are indistinguishable from one another in the Black Box study, and verification was not assessed.

Chapter 3 demonstrated there is substantial variability in the latent print evaluation, and Chapter 4
demonstrated the challenges in determining an expected answer for a given item. There is likely
examiner variation in interpretation of the analysis, comparison, and evaluation step. The best we can
do with the results from the Black Box study is quantify the examiner variation in the analysis step, and
variation in comparison and evaluation steps combined (i.e. the source decision). Additional studies
should be conducted to explicitly measure what examiners consider each of these steps to consist of,
and if there is variation in both definitions and performance at each step.

A better-defined procedure for latent print comparisons could lead to less variation as well as a better
understanding of where to focus training and development of tools to assist examiners.

If a procedure were defined in clearer terms, for instance:

1. Determine if latent print has value for comparison

2. Determine if there is enough overlapping area between latent and reference print to proceed

3. Locate minutiae on latent print

4. Locate minutiae on reference print

5. Decide if there is sufficient conflicting information to exclude
6. Decide if there is sufficient corresponding information to identify

Data could then be collected on each step of the process, and a straightforward IRTree could be constructed, with each point in the process (1-6) represented with a branch split.

Additionally, alternative methods of reporting exist (LR, Qualified Conclusions, etc.) (see Thompson (2018) for a discussion of presenting forensic source conclusions in general.) As evidenced by the FBI including the Close option as a reason for an inconclusive, the evidence must be quite strong in order to report an individualization. (Recall that the full text of a “close” response is: *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*). Under the current US standard for reporting (individualization, exclusion, or inconclusive), a ‘close’ counts as an inconclusive, even though it may contain far more information than an ‘insufficient’ inconclusive.

Under the current method of reporting, there is no way for the courts to use inconclusive determinations, rendering ‘close’ inconclusives useless. Results from Chapter 5 suggest that providing an option to report “soft” matches or non-matches may lead to a better understanding of an analyst’s evaluation of a pair of prints. Stoel (2017) suggested a likelihood-ratio based method of reporting, arguing that “when you allow for weaker conclusions, the justice system is stronger”. ‘Qualified Conclusion’ methods of reporting have also been advocated for in, e.g., Dror and Langenburg (2019).

Adopting a method of reporting based on a scale from ‘individualization’ to ‘exclusion’ would also allow for the separation of examiner proficiency (i.e. distinguishing matches from non-matches) from examiner “willingness to respond” (i.e. conclusive/inconclusive).

IRTrees for modeling examiner behavior would also be directly applicable to the likelihood ratio (LR) based method of reporting, which is commonly used throughout Europe and Canada. As an example, a score based LR may be represented as:

\[
LR = \frac{P(S = s| \text{Same Source})}{P(S = s| \text{Different Sources})}.
\]
where $S$ is a similarity score for a piece of evidence and a reference sample. LRs then exist on a continuous scale from $(0, \infty)$, with scores near 0 being supportive of an exclusion, and large scores being supportive of an individualization. Scores near one, however, do not provide support for either individualizations or exclusions and are therefore similar to an inconclusive. This suggests a natural transformation of an LR to a discretized bipolar response scale, which is represented in Figure 6.2. It is unclear how large or small a LR must be in order to provide a “conclusive” source evaluation, and those unknown points are represented with $X$ and $X'$ in the figure. Binning LR responses into $\text{Match (LR} > X\text{), Soft match (1 < LR < X), inconclusive(LR} \approx 1\text{), soft non-match (1 > LR > X')}$, and $\text{non-match}$ categories $(LR < X')$, would make such data directly applicable to the IRTrees framework introduced in this work.

![Figure 6.2: A visual description of how LR-based reporting methods could be adopted for the IRTrees framework.](image)

If data were collected in which analysts evaluated fingerprints on the likelihood ratio scale, we could not only learn more about examiner decision-making using a framework such as IRTrees, but also use such responses to evaluate different approaches to determining a likelihood ratio.

### 6.4.3 Broader Recommendations for Forensic Science

While the focus of this dissertation has been on latent print examination, this research could inform any area of forensic science in which subjective decision-making plays a role. While each forensic area uses a different standard for decision-making and reporting, IRT or IRTrees could be applied in a relatively straightforward way. As discussed in the previous section, the IRTrees framework is especially flexible for different reporting standards.
IRT should be adopted in all forensic proficiency testing. This would allow for the standardization of examiner scores across multiple years, adjusting for exams that were easier or harder than other exams. This would not only lead to a better understanding of examiner performance, but safeguard examiners if a given proficiency exam was too difficult.

Additionally, proficiency tests with a wider distribution of item difficulties should be adopted in order to provide a more precise estimate of examiner proficiency. Accurately measuring proficiency 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 proficiency tests could lead to targeted training for examiners in order to improve their proficiency. This is particularly true if multi-step proficiency tests were adopted, in which examiners were given feedback about where in the decision-making process they differed from other examiners.

While controlled proficiency test situations may not provide a full picture of performance in casework, results are nevertheless useful for assessing variation in examiner decisions. Furthermore, examiner error in controlled settings provides a lower bound for examiner error in uncontrolled settings (i.e. casework). In order to improve performance in casework, we must first be able to measure examiner performance and controlled proficiency test situations are a natural place to do so.

The National Institute of Justice and the National Institute of Standards and Technology convened an expert working group in response to the 2009 National Academies Report on Forensic Science [National Research Council, 2009]. One of the working group’s recommendations was “to create a culture in which both management and staff understand that openness about errors is not necessarily a path to punitive sanctions but rather is part of an effective system to detect deviations from desired practices and incorrect judgments in latent print casework” [NIST 2012]. A wide-scale adoption of IRT in forensic science could be one step towards creating such a culture in forensic science domains in which subjective decision-making plays a role.
Bibliography


President’s Council of Advisors on Science and Technology (2016). Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods. Technical report, Executive Office of The President’s Council of Advisors on Science and Technology, Washington DC.


143


