Law, Probability and Risk (2018) **17**, 111–121 doi:10.1093/lpr/mgy009
Advance Access publication on May 5, 2018

Proficiency testing of fingerprint examiners with Bayesian Item Response Theory[1]

Amanda S. Luby[†]

and

Joseph B. Kadane

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

[Received on 12 December 2017; revised on 13 March 2018; accepted on 24 March 2018]

In recent years, the forensic community has pushed to increase the scientific basis of forensic evidence, which has included proficiency testing for fingerprint analysts. We used proficiency testing data collected by Collaborative Testing Services in which 431 fingerprint analysts were asked to identify the source of latent prints. The data were analysed using a Rasch model with a Bayesian estimation approach. Although these data provide valuable information about the relative proficiency of the examiners and the relative difficulty of the questions, it does not necessarily extrapolate onto general performance of examiners or difficulty in casework, which we show through sensitivity analysis and simulation. We show that a Bayesian Item Response Theory (IRT) analysis provides a deeper understanding of analysts' proficiency and question difficulty than other forms of analysis. A large-scale adoption of IRT in this area would provide both more precise estimates of proficiency and quantitative evidence for the relative difficulty of different questions.

Keywords: fingerprint identification; reliability; proficiency testing; accuracy; Bayesian statistics; Rasch model; logistic regression.

1. Introduction

In many forensic science applications, proficiency tests are given to analysts with the results serving a variety of purposes, including training analysts, determining baseline competency levels, improving practices and procedures and identifying future needs (Koehler, 2013; AAAS, 2017). For instance, each examiner in a given laboratory may complete a proficiency exam each year in order to assess practices and performance. The proficiency exam may then be re-used for a new trainee to provide them with individual practice outside of casework. Proficiency exams may also be used to assess the reliability of available technology and determine whether new equipment is needed.

Proficiency tests are often high stakes for examiners, as passing a proficiency exam is used as evidence for an examiner's expertise in the courtroom. Laboratories may want to maintain a certain passing rate. Failing a proficiency exam could also jeopardize an analyst's career. If an exam contains

[†]Corresponding author: Email: aluby@andrew.cmu.edu

[1]The material presented here is based upon work supported in part under Award No. 70NANB15H176 from the U.S. Department of Commerce, National Institute of Science and Technology. Any opinions, findings, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Institute of Science and Technology, nor the Center for Statistics and Applications in Forensic Evidence.

an unusually hard question, the results could have far-reaching implications. Additionally, since many proficiency tests are open to anyone, there are likely to be large differences among test-takers when it comes to their proficiency and skill for latent print identification. We propose an analysis method that allows for quantification of individual differences in proficiency testing that goes beyond the raw score, and additionally accounts for varying difficulty of questions.

Item Response Theory (IRT) is often used in educational testing to account for differences between test-takers and differences in test question difficulty when analysing exams (Rasch, 1960; Lord, 1980; de Boeck and Wilson, 2004; Fischer and Molenaar, 2012). This approach allows for measurement of not only the performance of individual test-takers, but also the difficulty of individual questions. Furthermore, IRT allows participants to be compared on the same scale, even if they were shown a different set of questions. Kerkhoff *et al.* (2015) have discussed the use of IRT for analysing firearm proficiency test results, but Item Response Models have yet to be implemented on forensic proficiency exams. Conclusions from a model of this type allow for a better understanding of how examiner proficiency and question difficulty contribute to the observed test responses in a forensic proficiency exam setting.

We propose the adoption of IRT in forensic proficiency settings in order to better understand the proficiency of examiners as well as the difficulty of questions and exams. If an IRT approach is adopted across many different tests, more precise estimates of proficiency would be possible. Item Response Models could also provide quantitative evidence for the relative difficulty of different questions across different exams.

2. Data

Collaborative Testing Services, Inc (CTS) provides interlaboratory proficiency tests that also meet requirements for external accreditation. Although CTS provides proficiency tests for a variety of forensic disciplines including: forensic biology/DNA, drug analysis, firearms and toolmarks, latent print and impressions, documents, trace evidence, toxicology, crime scene and digital and multimedia evidence, we choose to focus on the results from a latent fingerprint proficiency test.

In particular, we analysed *Latent Print Examination Test Number 16-515/516*,[2] which included results from 431 respondents and consisted of 12 identification questions. Participants were given four possible known donors (denoted A, B, C and D) with each palm print in addition to the full set of 10 fingerprints. They were also given 12 fingerprints of unknown source taken from a hypothetical crime scene, and were asked to identify the donor and finger of each of the fingerprints. They were also given the option to not identify ('NI') each print.

As noted in the published report from CTS, '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 testing environment is not controlled, it is unclear whether each response corresponds to an individual examiner or represents the consensus answer of a group of examiners working together on the exam. We will use 'participant', 'examinee' or 'respondent' to denote a set of responses to the exam. Results describe either the individual or group that responded to the exam.

[2]See http://www.ctsforensics.com/assets/news/3616_Web.pdf.

The data are provided as a PDF table in which incorrect answers are identified. The majority of participants ($n = 383$) correctly answered all 12 of the questions. Eleven of the 12 questions had over 98% correct response rates, with one question (Q2) having a 100% correct response rate.

3. IRT framework

IRT is used extensively in psychometrics and educational testing theory to study the relationship between a respondent's (unobserved) proficiency and their performance on a certain item. For analysing test results using an IRT approach, the probability of a correct response to a question depends on both (a) the difficulty of the question and (b) the proficiency of the respondent (Fischer and Molenaar, 2012). The proficiency and difficulty parameters are assumed to measure the same underlying latent trait, in this case, proficiency in latent print identification tasks. Both 'difficulty' and 'proficiency', however, are not observable and so they must be estimated using the available data. IRT allows for participants to be compared on the same scale, even if they were shown different sets of questions in an exam. Item Response Models can account for varying difficulty of questions, and adjust each individual's proficiency estimate accordingly.

The model we implement incorporates a parameter for each individual who took the exam, and a single parameter for each item. This two-parameter logistic model is known as the Rasch model (Rasch, 1960). Each of the N people who took a given exam are associated with a 'proficiency', θ_n, and each of the I items is associated with a 'difficulty', b_i.

The data are represented as a matrix of binary responses, where 1 indicates a correct answer and 0 represents an incorrect answer. In the context of proficiency exams, if we have n participants and m different questions, we can then express the data as a $n \times m$ matrix of participant responses to the exam:

$$Y = \begin{bmatrix} 1 & - & - & \dots & 0 \\ - & 0 & - & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ - & 1 & 1 & \dots & - \end{bmatrix},$$

where $Y_{pi} = 1$ (row p, column i) if participant p correctly answered question i, $Y_{pi} = 0$ if participant p incorrectly answered question i, and $Y_{pi} = -$ indicates that participant p was not shown question i. For the CTS proficiency test, each participant is shown every question, but the model framework is also appropriate for analysing tests in which participants were shown different sets of questions.

The probability of observing a correct response on a question is then:

$$P(X_{ni} = 1) = \frac{\exp\,(\theta_n - b_i)}{1 + \exp\,(\theta_n - b_i)}. \tag{1}$$

The model, as written, is not identified. That is, for any two estimates of θ_n and b_i, α can be subtracted from θ_n and added to b_i, and produce the same probability. For this reason, there are often constraints placed on the parameters such as forcing some question to have difficulty equal to zero; or requiring the sum of the difficulties to be equal to some constant.

Given the model above, when the test-taker's proficiency, θ_n, is equal to an item's difficulty, b_i, the probability of person n correctly answering question i is equal to 0.5. If the test-taker's proficiency is larger than the question difficulty, the probability of a correct response is closer to 1. If the test-taker's proficiency is smaller than the question difficulty, the probability of a correct response is closer to 0.

The Rasch model requires estimation of $I + N$ parameters, one for each participant and question. We discuss two perspectives for estimating these parameters, a maximum likelihood estimation approach and a Bayesian estimation approach.

3.1 *Maximum likelihood estimation*

To obtain maximum likelihood estimates for both the difficulty and proficiency parameters, a joint maximum likelihood estimation (JMLE) procedure is used. In tests with few items, the JMLE estimates for the proficiency parameters are biased, which also leads to misestimation of difficulty parameters due to the iterative estimation procedure (Lord, 1986). Using a JMLE estimation procedure also leads to infinite estimates for perfect or zero scores, which is undesirable for a proficiency exam and is illustrated below.

The likelihood for the model given in (1) is:

$$L(\theta, \mathbf{b}|X) = \frac{\Pi_N \Pi_I \exp\left(x_{ni}(\theta_n - b_i)\right)}{\Pi_N \Pi_I 1 + \exp\left(\theta_n - b_i\right)}. \tag{2}$$

The maximum likelihood estimates for θ_n and b_i then satisfy

$$r_n = \sum_I \frac{\exp\left(\theta_n - b_i\right)}{(1 + \exp\left(\theta_n - b_i\right))}, n = 1, \ldots, N \tag{3}$$

and

$$s_i = \sum_N \frac{\exp\left(\theta_n - b_i\right)}{(1 + \exp\left(\theta_n - b_i\right))}, i = 1, \ldots, I, \tag{4}$$

where r_n and s_i are the marginal totals for each person and item, respectively, and are the sufficient statistics for θ_n and b_i.

For this CTS exam in particular, all 431 participants correctly answered Question 2. The maximum likelihood estimates for θ_n and b_2 would then need to satisfy:

$$s_2 = \sum_N \frac{\exp\left(\theta_n - b_2\right)}{(1 + \exp\left(\theta_n - b_2\right))}$$

$$431 = \sum_1^{431} \frac{\exp\left(\theta_n - b_2\right)}{(1 + \exp\left(\theta_n - b_2\right))},$$

for which a solution does not exist. This happens whenever $r_n = 0$ or I or $s_i = 0$ or N. In the broader context of proficiency exams, some examinees would likely score 100% on most exams, and it is certainly possible that an examinee would not answer any questions correctly. In a maximum likelihood estimation framework, adaptations such as *Conditional Maximum Likelihood Estimation*

(Andersen, 1970) and *Marginal Maximum Likelihood Estimation* (Bock and Aitkin, 1981) are often used in practice to account for this issue, each with their own features and drawbacks.

3.2 *Bayesian estimation*

An alternative approach to a maximum likelihood procedure is implementing a Bayesian estimation approach and incorporating prior distributions for each of the parameters of interest. We proceed with such an approach, and assume each θ_n and b_i have independent normal distributions,

$$\theta_n \sim N(0, \sigma_\theta^2) \text{ and } b_i \sim N(\mu_b, \sigma_b^2),$$

with the same logistic probability model as in (1). Note that the prior expectation of the proficiency parameters, θ, is 0 in order to fix the origin of the latent scale. Thus the difficulty parameters, b_i, are located relative to the average proficiency level in the population. After estimating the item difficulties and participant abilities, the questions that are more difficult (relative to the average proficiency) will have a positive difficulty estimate, while the easier questions will have a negative difficulty estimate. Difficulties and abilities are estimated using a Markov Chain Monte Carlo (MCMC) Gibbs sampler.

4. Results

We tested four combinations of normal prior distributions:

(1) $\sigma_\theta^2 = 1,\ \mu_b = 0,\ \sigma_b^2 = 1$
(2) $\sigma_\theta^2 = 1,\ \mu_b \sim N(0, 1000),\ \sigma_b^2 = 1$
(3) $\sigma_\theta^2 = 1000,\ \mu_b = 0,\ \sigma_b^2 = 1000$
(4) $\sigma_\theta^2 = 1000,\ \mu_b \sim N(0, 1000),\ \sigma_b^2 = 1000.$

Due to the lack of variation in the data, models in which large variances were more likely (3 and 4) led to poor convergence. The estimates (posterior means) for each of the question difficulty parameters, along with the posterior standard deviation, are given below in Table 1.

Models 1 and 2 produced similar estimates, as did Models 3 and 4. Allowing the mean of the item difficulties to 'float', in this case, does not have a large impact on the posterior distributions. There is, however, a large difference between the first two models and the second two models due to the change in item difficulty variance (σ_b^2). Interestingly, there is a slightly different ordering of the easiest questions in the test (Q2, Q4, Q7, Q11) for each of the four models, although the differences in estimates for the four questions are quite small.

The change in prior variance in the item difficulties has a noticeable impact on the proficiency estimates as well, as evidenced by the person proficiency posterior distributions in Fig. 1. In Models 1 and 2, there are distinct distributions for each of the observed scores. In Models 3 and 4, in which a larger variance was used, clusters of distributions are not distinguishable, except for a small cluster centred near -25 corresponding to the lowest proficiency participants.

4.1 *Coding correct responses*

Proceeding with the results from Model 1 for illustration, the proficiency posteriors in Fig. 2 show a similar distribution, with a mean slightly above 0, was estimated for most of the people who took the exam. There is a fairly large cluster of distributions with a mean slightly below 0, which correspond to

TABLE 1 Posterior means and standard deviations from MCMC samples

	Model 1		Model 2		Model 3		Model 4	
	Mean	sd	Mean	sd	Mean	sd	Mean	sd
Q1	−4.45	0.39	−5.60	0.55	−24.49	8.88	−25.42	9.35
Q2	−4.90	0.42	−6.80	0.78	−26.48	9.39	−27.39	9.77
Q3	−4.60	0.41	−5.89	0.63	−25.02	8.97	−25.77	9.37
Q4	−4.75	0.42	−6.24	0.71	−25.36	8.97	−26.25	9.66
Q5	−4.21	0.34	−5.12	0.45	−23.91	8.81	−24.82	9.26
Q6	−2.43	0.18	−2.76	0.19	−19.79	8.26	−20.67	8.71
Q7	−4.72	0.43	−6.20	0.66	−25.70	9.30	−26.52	9.69
Q8	−4.22	0.34	−5.15	0.45	−23.93	8.76	−24.85	9.30
Q9	−4.09	0.33	−4.95	0.44	−23.71	8.77	−24.58	9.26
Q10	−4.26	0.34	−5.32	0.50	−24.15	8.83	−25.12	9.38
Q11	−4.70	0.44	−6.21	0.76	−25.32	9.10	−26.30	9.60
Q12	−4.57	0.41	−5.87	0.65	−24.90	8.89	−25.80	9.40

the estimates for examiners who only answered one question incorrectly. The remaining distributions correspond to examiners who incorrectly answered more questions.

The logistic curves in Fig. 3, show that one of the questions (Q6) stands out from the rest of the questions on the exam. The consensus answer for question six was the right index finger of person D (denoted 'D, RI'). Out of the 31 people who incorrectly answered this question, 23 of them answered with the palm print of person D (denoted 'D, RP'). Since these participants matched the consensus answer for the donor of the print (D), as well as the hand (right), scoring their responses as incorrect may not reflect their actual performance.

Examining the comments from participants provides insight into this pattern. Many of the comments refer to question six, with participants making notes such as 'Q-6, I used Item D right palm print exemplar to identify the right index finger phalange', 'Q6 labelled RP originally, changed to RI 2nd joint. Photo taken from RP page not FP page' and 'Q6 identified to the 2nd & 3rd joints of finger #2 (RI)'. At least nine of the examiners who got the question wrong did correctly identify the right index finger of person D, but used the palm print exemplars rather than the individual exemplar print, which was reflected in their response.

This leads to a question of which answers should count as correct in this case. Is a response 'correct' if the correct person is identified, or does it need to be the correct finger? If the crime scene sample print was actually not from the right index fingerprint, but from a different part of the finger that was visible only on the reference palm print, should the right index response then be counted as incorrect?

If one considers 'D, RP' as the only correct response, the logistic curves shown in Fig. 4 result. Question 6 is very far away from the other curves, with a much higher θ needed to increase the probability of correctly answering the question.

Perhaps it makes more sense to consider both 'D, RI' and 'D, RP' as correct. Based on the comments from participants, many either used the right palm image to identify the right index, or reported the right palm and noted a mark on the right index as an identifier, so considering both responses to be correct is not unreasonable. Scoring the exam in this way leads to the logistic curves shown in Fig. 5, with Question 6 closer to the rest of the questions than in Figs 3 or 4, but still noticeably different.

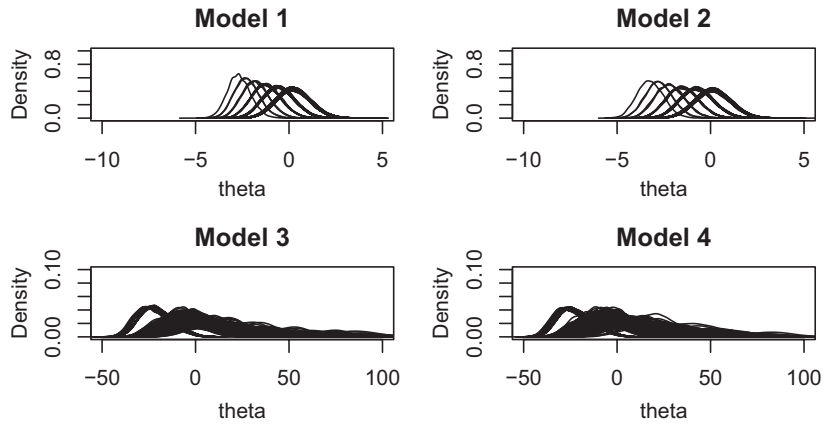

FIG. 1. Posterior densities for the person parameters, θ, for each of the four models.

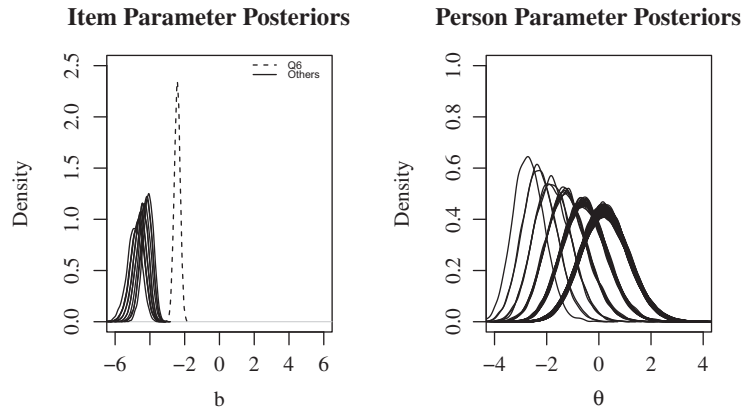

FIG. 2. Posterior distributions for item parameters (left) and person parameters (right).

4.2 *A fictional test*

The Rasch model is sensitive to prior distribution choice, as shown at the beginning of this section. Section 4.1 illustrates the sensitivity to scoring protocol. Sensitivity to question difficulties is also important. Because the CTS test examined here lacks a variety of question difficulties, a simulation is used to address sensitivity to question difficulty.

Suppose the same test is given to a group of 200 people consisting of 100 examiners and 100 novices. Assume that the proficiency for the examiners is close to the posterior mean of the group that correctly answered every question, $\theta_n = 0.25$. Also assume that the novices are more proficient than the lowest scorers, but are not as proficient as the highest scores, and set their proficiency level to $\theta_n = -1$, which is near the posterior mean of all participants who did not get all 12 questions correct. Also assume each of the b_is is equal to their posterior mean from Model 1 (Table 1).

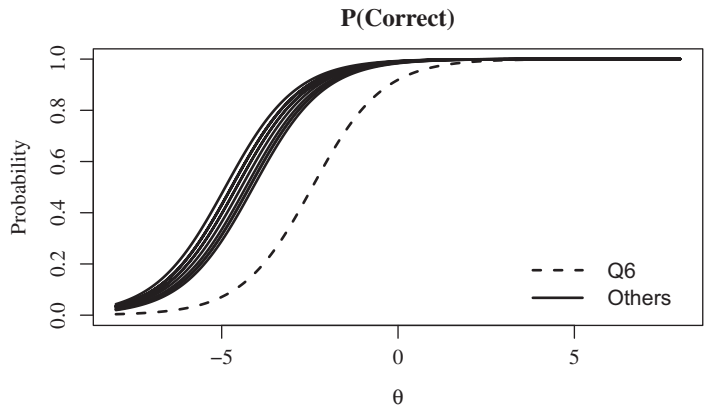

FIG. 3. Logistic curves for the probability of observing a correct response at different proficiency levels, for each question. Question 6 is scored with 'D, RI' as correct.

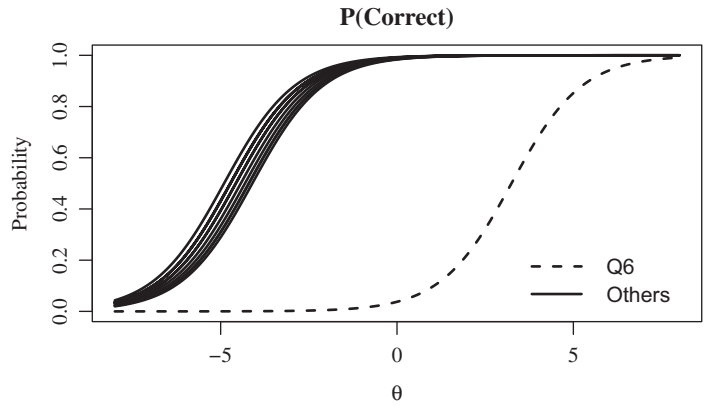

FIG. 4. Logistic curves for the probability of observing a correct response at different proficiency levels, for each question. Question 6 is scored with only 'D, RP' as correct.

Simulating new test results using the probabilities from the logistic model leads to the results in Table 2. Although novices score slightly lower on average than the experts, discrimination between novices and experts is not possible at an individual level. That is, if someone received a perfect score, it is not clear whether they are an examiner or a novice, since both novices and examiners mostly answered 12 questions correctly.

If the simulation is repeated with a wider range of b_is, which would correspond to a test with both easier and harder questions, rather than a test with all of the questions at a similar difficulty, different results are obtained. Parameter values of $\theta_n = -1$ and 0.25 for novices and examiners, respectively, are kept from the previous simulation, and the difficulties (b_i) are drawn from a $N(0, 1)$ distribution. The process of simulating correct answers based on the logistic probabilities from the model is repeated, yielding the results in Table 3.

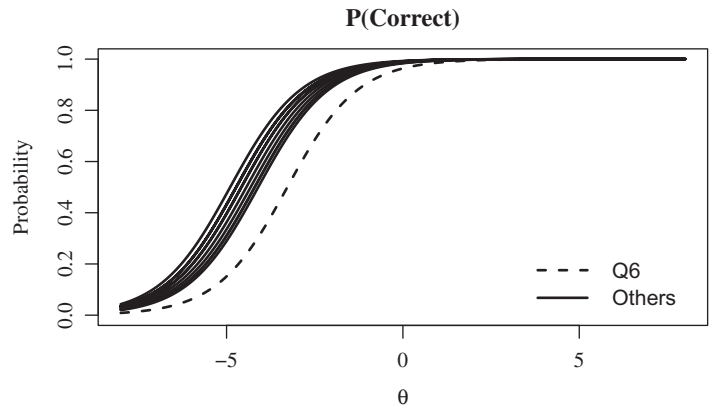

FIG. 5. Logistic curves for the probability of observing a correct response at different proficiency levels, for each question. Question 6 is scored with either 'D, RI' or 'D, RP' as correct.

TABLE 2 Results of Simulation 1

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Novices	10	11	12	11.48	12	12
Examiners	10	12	12	11.83	12	12

This simulation results in a larger difference between novices and examiners, as well as discrimination between the two groups at an individual level. For instance, if five were set as the 'passing threshold' for the exam, less than 1/4 of the examiners would fail, while only about 1/4 of novices would pass. The difficulties of the questions could, of course, be adjusted to make this overlap of novices and examiners as large or small as desired.

5. Discussion and future work

Posterior analysis of the data under the Rasch model shows that there is one ambiguous question (Question 6) in this test, which was estimated to have a higher difficulty than the others, but most questions were estimated to be of lower difficulty and quite similar to one another. The person parameters behave as expected given the item parameters: those who answered easier questions incorrectly are given lower estimates of proficiency.

A Rasch model allows the effects of participants and questions to be evaluated in conjunction with one another. Implementing a Bayesian approach provides a natural framework to solve problems with estimation in cases where either a participant gets every question correct or incorrect, or a question that every participant answers correctly or incorrectly. The model is sensitive to prior distribution choice, alternative scoring mechanisms and differences in question difficulty. Simulations allow for illustrations of what may happen if a test is designed in a certain way and given to a certain group of people.

Simulation results suggest that tests with a greater range of difficulty of the questions would be more informative when estimating examiner proficiency. Tests that include harder questions would

TABLE 3 *Results of Simulation 2*

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Novices	0	2	4	3.57	5	9
Examiners	3	5	7	6.55	8	10

discriminate between different levels of proficiency, as in our fictional examiners and novices, but also would allow examiners to better understand their proficiency relative to their peers in more difficult comparison tasks.

As mentioned in Section 2, aspects of the testing environment may influence the results of the exam, such as multiple participants coming to a consensus answer or access to additional tools, but differences in test administration are often not reported for forensic proficiency exams and there is no way to measure or control for these differences. Although standardizing testing procedures across all forensic laboratories may be infeasible, if additional data were to be collected, Item Response Models could be adapted to account for differences in testing environment among participants.

A common concern with forensic proficiency tests, as they are currently designed, is whether accuracy rates in casework can be inferred from the results (Koehler, 2013; AAAS, 2017). Although this work focused on the traditional class of proficiency tests, a Rasch model (or more complex IRT models) could be used on tests designed for the purpose of assessing accuracy rates (such as Ulery *et al.* (2011)), to further understand how the individual differences among examiners and questions contribute to the results.

This modelling framework is very flexible, and would allow for a natural hierarchical structure for both person and item parameters (de Boeck and Wilson, 2004). We could, for instance, model the person abilities based on where they were trained, how many years of experience they have, where they work now and other individual attributes. We could also model the difficulties of questions based on, for instance, whether the prints were taken from a real crime scene or are synthetic, if the matches were found using an Automated Fingerprint Identification System (AFIS) search or a known suspect, and if the print is partial or complete. Using a hierarchical model of this nature would show whether there are some types of questions that some types of examiners are more likely to get wrong, and the results could be used to provide targeted training based on those questions.

A further extension to the Rasch model is to include additional latent (unobservable) variables for the questions. The two-parameter logistic model includes a discrimination parameter for each of the questions, and the three-parameter logistic model includes a pseudo-guessing parameter for each of the questions (Lord, 1980; Harris, 1989). The additional parameters increase the amount of computational complexity needed for estimation. There is not enough variation in the particular CTS data used in Section 4 to warrant the use of such models, but a large-scale adoption of IRT models to forensic proficiency data may require the use of a more complex model.

REFERENCES

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

Andersen, E. B. (1970), 'Asymptotic properties of conditional maximum-likelihood estimators,' *Journal of the Royal Statistical Society. Series B (Methodological)*, **32**(2), 283–301.
Bock, R. D., and Aitkin, M. (1981), 'Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm,' *Psychometrika*, **46**(4), 443–459.
de Boeck, P., and Wilson, M. (2004), *Explanatory Item Response Models: A generalized linear and nonlinear approach*, New York: Springer.
Fischer, G. H., and Molenaar, I. W. (2012), *Rasch models: Foundations, recent developments, and applications*, New York: Springer Science & Business Media.
Harris, D. (1989), 'Comparison of 1-, 2-, and 3-Parameter IRT Models,' *Educational Measurement: Issues and Practice*, **8**(1), 35–41.
Kerkhoff, W., Stoel, R., Berger, C., Mattijssen, E., Hermsen, R., Smits, N., and Hardy, H. (2015), 'Design and results of an exploratory double blind testing program in firearms examination,' *Science & Justice*, **55**(6), 514–519.
Koehler, J. J. (2013), 'Proficiency tests to estimate error rates in the forensic sciences,' *Law, Probability and Risk*, **12**(1), 89–98.
Lord, F. M. (1980), *Applications of item response theory to practical testing problems*, Hillsdale, NJ: Erlbaum.
Lord, F. M. (1986), 'Maximum likelihood and Bayesian parameter estimation in item response theory,' *Journal of Educational Measurement*, **23**(2), 157–162.
Rasch, G. (1960), *Probabilistic models for some intelligence and attainment tests*, Chicago: University of Chicago Press.
Ulery, B. T., Hicklin, R. A., Buscaglia, J., and Roberts, M. A. (2011), 'Accuracy and reliability of forensic latent fingerprint decisions,' *Proceedings of the National Academy of Sciences of the USA*, **108**(19), 7733–7738.