Educational and Psychological
Measurement
2022, Vol. 82(5) 989–1019

DOI: 10.1177/00131644211046253
journals.sagepub.com/home/epm

Non-iterative Conditional Pairwise Estimation for the Rating Scale Model

Mark Elliott[1] and Paula Buttery[1]

Abstract

We investigate two non-iterative estimation procedures for Rasch models, the pairwise estimation procedure (PAIR) and the Eigenvector method (EVM), and identify theoretical issues with EVM for rating scale model (RSM) threshold estimation. We develop a new procedure to resolve these issues—the conditional pairwise adjacent thresholds procedure (CPAT)—and test the methods using a large number of simulated datasets to compare the estimates against known generating parameters. We find support for our hypotheses, in particular that EVM threshold estimates suffer from theoretical issues which lead to biased estimates and that CPAT represents a means of resolving these issues. These findings are both statistically significant ($p < .001$) and of a large effect size. We conclude that CPAT deserves serious consideration as a conditional, computationally efficient approach to Rasch parameter estimation for the RSM. CPAT has particular potential for use in contexts where computational load may be an issue, such as systems with multiple online algorithms and large test banks with sparse data designs.

Keywords

Rasch, rating scale model, estimation, pairwise, simulation

Introduction

While modern computing speeds mean that most estimation procedures are fast enough for a standard, offline Rasch analysis to be run, there are contexts where computational expense is an important consideration, for example larger scale

[1] University of Cambridge, Cambridge, UK

Corresponding Author:
Mark Elliott, University of Cambridge, William Gates Building, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK.
Email: mwe24@cam.ac.uk

automated systems where repeated calibration runs may be required alongside other processes, and large adaptive testing banks with hundreds or even thousands of items. For such contexts, some procedures such as conditional maximum likelihood estimation (CMLE) (Andersen, 1970, 1973) may still prove too computationally expensive for practical use. In this context, non-iterative procedures, which do not require repeated runs, possess an advantage.

Another desirable property of a Rasch estimation procedure is that it takes advantage of the Rasch property of the existence of sufficient statistics to condition out person parameters (Andrich & Luo, 2003), resulting in consistent estimates (Andersen, 1973), unlike the unconditional joint maximum likelihood estimation procedure (JMLE; Wright & Masters, 1982; Wright & Panchapakesan, 1969), which simultaneously estimates person and item parameters and suffers from the incidental parameter problem (Ghosh, 1994; Lancaster, 2000; Neyman & Scott, 1948) as a result, since the number of parameters to be estimated increases with the sample size. Conditional estimation methods also avoid the need for distributional assumptions as with marginal approaches, which can lead to estimation bias where person abilities are not normally distributed (Zwinderman & van den Wollenberg, 1990). The only established methods which are both conditional and non-iterative are PAIR (Choppin, 1968, 1985; Wright & Masters, 1982), a procedure based on pairwise comparisons of response data for dichotomous items, and the closely related eigenvector method (EVM; Garner & Engelhard, 2009; Garner, 2002), which also extends the PAIR approach fully to the rating scale model (RSM; Andrich, 1978) and many-facet Rasch models (MFRM; Linacre, 1994).

Our focus in this study is on the RSM; we will begin by reviewing the formulation of the RSM, PAIR, and EVM, before discussing theoretical issues, proposing solutions and formulating and testing hypotheses using simulated datasets.

Rating Scale Model

The RSM is a specific formulation of a polytomous Rasch model suitable for tests which are marked on an ordinal scale with three or more categories according to a set of rating criteria (items henceforth), for example an essay marked on a 0 to 5 scale for grammar, vocabulary, organization, and content. The model is also used for questionnaires using Likert scales. The RSM is formally identical to the partial credit model (PCM; Masters, 1982), except that the threshold structure is constrained to be the same across all items. Mathematically, the RSM formulation giving the probability P_{nik} of person n achieving a score of k on item i with a shared maximum available score of m is given, in probability ratio form, by:

$$\ln\left(\frac{P_{nik}}{P_{ni(k-1)}}\right)=\beta_n-\delta_i-\tau_k \tag{1}$$

As Equation 1 indicates, the overall item location parameter $\{\delta_i\}$ and the thresholds $\{\tau_k\}$ are independent of each other.

PAIR

Choppin (1968, 1985) developed the conditional pairwise PAIR algorithm, focusing on the dichotomous Rasch model, or simple logistic model (SLM; Rasch, 1960, 1961). The SLM is a further restricted form of the RSM, with one threshold (with location 0) for each item, with the corresponding simplified equation:

$$\ln\left(\frac{P_{ni1}}{P_{ni0}}\right)=\beta_n-\delta_i \tag{2}$$

If we consider a person n responding to a pair of items (i, j), Equation 2 gives:

$$\ln\left(\frac{P_{ni1}}{P_{ni0}}\frac{P_{nj0}}{P_{nj1}}\right)=(\beta_n-\delta_i)-(\beta_n-\delta_j)=\delta_j-\delta_i \tag{3}$$

Assuming that the requirement for local item independence is met, this is equivalent to:

$$\ln\left(\frac{P_{ni1\wedge nj0}}{P_{ni0\wedge nj1}}\right)=\delta_j-\delta_i \tag{4}$$

This result means that, in the case where a person responds to only one of a pair of items (i, j) correctly, the probability that the correctly answered item is i or j is a function only of the difference in the difficulty of the two items: person ability has been conditioned out.

Choppin (1968, 1985) developed methods to produce a vector of item difficulties from Equation 4 in both iterative and non-iterative ways. The non-iterative method, PAIR, is based on substituting the observed counts of $ni1\wedge nj0$ and $ni0\wedge nj1$ for $P_{ni1\wedge nj0}$ and $P_{ni0\wedge nj1}$ in Equation 4 to provide an estimate of $e^{\delta_j-\delta_i}$ for each pair of items (i,j) in a process involving several steps:

1. Create a matrix C of conditional category response frequencies $ni1\wedge nj0$ for all possible pairs of items (i, j).
2. From this matrix C, dividing C_{ij} by C_{ji} provides an estimate of $e^{\delta_j-\delta_i}$ based on the responses to items i and j. However, C contains zero values – the leading diagonal is comprised of zeros and there may also be cases where $ni1\wedge nj0$ returns a zero value. Choppin (1985) showed algebraically that raising C to successive integer powers will eventually remove all zero values provided C is well-conditioned (Fischer, 1981), and that the underlying structure of the resulting matrix is maintained—in effect, direct comparisons between a pair of items i and j are supplemented by additional comparisons mediated by another item k in the case of C^2, and further mediations are added with successive powers.
3. From the resulting matrix C^n, generate a new matrix D such that $D_{ij}=C^n_{ji}/C^n_{ij}$. D is a positive pairwise reciprocal matrix: $D_{ij}=1/D_{ji}$, where $D_{ij}>0$. Matrix

D is a matrix of pairwise comparisons of item difficulties once the person parameters have been conditioned out, which is identical to the pairwise comparison model developed by Zermelo (1929) to determine relative player strength from a chess tournament where there may be missing data, the values in the columns of which relate to item difficulties on the ratio scale form of the Rasch model. Bradley and Terry (1952) independently developed an interval scale version of the same model, which is directly equivalent to the standard log-odds form of the Rasch model with person parameters conditioned out.

4. From the matrix D, take the arithmetic mean of each row to create a vector of item difficulty estimates; these represent the item difficulty estimates on the ratio scale version of the Rasch model (Rasch, 1960).
5. To convert to the interval scale Rasch model, take the natural logarithm of each element of the difficulty vector and, for convention, subtract the mean difficulty so that $\sum_i \delta_i = 0$.

The primary motivation for Choppin's (1985) work was the sensitivity of previous methods to missing data, and a weakness which extends to CML: since CML is predicated on the probability of complete response vectors across a set of items, and the studies have indicated that it may be sensitive to missing data (Eggen & Verhelst, 2006; Heine & Tarnai, 2015), whereas PAIR is derived purely from the minimal data required to compare a pair of items; missing data only has an effect to the extent that it reduces the number of comparisons available between a pair of items. Heine and Tarnai (2015) compare the estimates (using the PCM) with randomly removed data to the estimates obtained from complete datasets using PAIR, CML, and MML. Heine and Tarnai find that PAIR outperforms CML, which is often regarded as the gold standard of estimation methods, in terms of the stability of the resulting estimates; PAIR and MML were similar in terms of stability. One limitation of Heine and Tarnai's findings is that, since they are based on authentic data, there is no definitive ground truth of known difficulty values against which to determine performance.

For polytomous responses, PAIR can be extended to the PCM (Masters, 1982) by treating each item threshold as an individual item and counting the conditional category frequencies for pairs of thresholds on different items where persons have scored in the adjacent categories either side of the two thresholds (Garner, 2002). It cannot, however, be used in the same manner for the RSM (Andrich, 1978), due to the constraints imposed on the threshold structure, which are identical across items; an alternative formulation is required.

The EVM

Mainly following Choppin's approach, (Garner 2002) and Garner and Engelhard (2009) apply an approach to the derivation of the item difficulty vector which differs only in the averaging, step, using the eigenvector corresponding to the principal eigenvalue of the matrix *D* rather than the arithmetic means of the rows. This approach, the EVM, has its foundation in the analytical hierarchy process (AHP; Saaty, 1994; Saaty & Bennett, 1977). Garner and Engelhard also extend the applicability of EVM from the SLM and the PCM to the RSM (Garner 2002) and further (Garner and Engelhard 2009) to the MFRM (Linacre, 1994).

For the RSM, we can condition out the person ability and Rasch-Andrich thresholds to obtain a formulation for the estimation of the overall item parameters by considering probabilities of scoring 1 more on item *i* than on item *j* or 1 more on item *j* than on item *i* conditional on the scores being in adjacent categories:

$$\ln\left(\frac{P_{nik}}{P_{ni(k-1)}}\frac{P_{nj(k-1)}}{P_{njk}}\right)=(\beta_n-\delta_i-\tau_k)-(\beta_n-\delta_j-\tau_k)=\delta_j-\delta_j \tag{5}$$

or, using local item independence:

$$\ln\left(\frac{P_{nik\wedge nj(k-1)}}{P_{ni(k-1)\wedge njk}}\right)=\delta_j-\delta_i \tag{6}$$

Garner and Engelhard go on to provide a formulation for the estimation of the Rasch–Andrich thresholds:

$$\ln\left(\frac{P_{nik}}{P_{ni(k-1)}}\frac{P_{ni(l-1)}}{P_{nil}}\right)=(\beta_n-\delta_i-\tau_k)-(\beta_n-\delta_i-\tau_l)=\tau_l-\tau_k \tag{7}$$

or, again using local item independence and taking exponents:

$$\ln\left(\frac{P_{nik\wedge ni(l-1)}}{P_{ni(k-1)\wedge nil}}\right)=\tau_l-\tau_k \tag{8}$$

Garner's (2002) approach to estimating this is to compare the number of pairs of persons with the same total score who scored *k* and $l-1$ on item *i* to the number of pairs of persons who scored $k-1$ and *l* on the same item (p. 115). This assumes the presence of a complete data matrix with no missing data, since the same total score on a different number of items will correspond to a different ability estimate. In order to extend their approach to datasets with missing data, it is necessary to subset the sample into sets of persons who responded to the same items; pairs would then be then counted within each subset and summed across all subsets.

There are certain theoretical issues caused by features of this approach to generating threshold estimates, which conditions out both the item difficulties and person abilities simultaneously in order to generate the conditional category frequency matrix for the thresholds. Firstly, for estimation purposes the person parameters cannot be said to have been fully conditioned out, since persons are grouped according to the same total score, meaning that person ability must be taken into account during the estimation procedure; this may result in biased estimates due to the incidental parameter problem (Ghosh, 1994; Lancaster, 2000; Neyman & Scott, 1948). Secondly, while two persons with the same total score on the same set of items will obtain the same ability estimate due to the sufficiency of the raw score statistic for person parameter estimation, this does not mean that they have the same true ability—the model considers the raw score to the outcome of a stochastic process mapping a continuous variable (ability on the latent trait) onto a closed set of discrete points corresponding to the set of possible raw scores. In reality, the only person who can be said with certainty to have the same underlying ability as a person is that same person, and the use of estimates as a proxy for true ability introduces a source of error. Finally, where there is missing data, the dataset must be subsetted further according to items responded to, since ability scores on the same number of responses to different subsets of items will have different ability estimates and therefore cannot be grouped together. This means more responses are unused, which will be detrimental to estimation accuracy. In fact, all persons who have a unique combination of scores and response patterns are unusable, while others may have small numbers of persons with whom comparisons can be made. For these reasons, we can hypothesize that EVM threshold estimates are likely to be more sensitive to missing data than item difficulty estimates, which do not suffer from the same issues since they follow the PAIR approach apart from the averaging step. The subsetting process also adds computational expense.

These observations, particularly the third point, indicate that the a theoretically correct approach to pairwise estimation should consider the scores obtained by the same person on two different items, that is, by counting comparisons of pairs of columns, as it is the case for item difficulty estimates—a modified approach is required.

Although it is not possible to condition out both person and item parameters simultaneously while counting columns, it is straightforward to condition out the person ability parameter:

$$\ln\left(\frac{P_{nik}}{P_{ni(k-1)}}\frac{P_{nj(l-1)}}{P_{njl}}\right) = (\beta_n - \delta_i - \tau_k) - (\beta_n - \delta_j - \tau_l) = (\delta_j - \delta_i) + (\tau_l - \tau_k) \tag{9}$$

or, using local item independence and re-arranging terms:

$$\tau_l - \tau_k = \ln\left(\frac{P_{nik \wedge nj(l-1)}}{P_{ni(k-1) \wedge njl}}\right) + (\delta_i - \delta_j) \tag{10}$$

Equation 10 would constitute a sufficient statistic, except that it contains the difference between the difficulties, $\delta_j - \delta_i$. However, this can be eliminated by considering the item pair (j, i) instead of (i, j); reversing i and j in Equation 10 while keeping thresholds k and l in the same order gives:

$$\tau_l - \tau_k = \ln\left(\frac{P_{ni(l-1) \wedge njk}}{P_{nil \wedge nj(k-1)}}\right) + (\delta_j - \delta_i) \tag{11}$$

Combining Equations 10 and 11 gives:

$$\tau_l - \tau_k = \frac{1}{2}(\ln\frac{P_{nik \wedge nj(l-1)}}{P_{ni(k-1) \wedge njl)}} + \ln\frac{P_{ni(l-1) \wedge njk}}{P_{nil \wedge nj(k-1)}}) \tag{12}$$

Using Equation 12, it is possible to construct an $m \times m$ matrix T of threshold differences with $T_{ij} = \tau_i - \tau_j$. This matrix is the equivalent of the element-wise natural logarithm of the reciprocal pairwise matrix in PAIR/EVM. Taking the arithmetic mean of row T_i:

$$\overline{T}_i = \frac{1}{m}\sum_{j=1}^{m} T_{ij} = \frac{1}{m}\sum_{j=1}^{m}(\tau_i - \tau_j) = \frac{1}{m}\sum_{j=1}^{m}\tau_i - \frac{1}{m}\sum_{j=1}^{m}\tau_j = \tau_i \tag{13}$$

due to the identity $\sum_{j=1}^{m}\tau_j = 0$. This means that the vector of arithmetic row means of T is a set of threshold estimates for thresholds τ_k, $k \in \{1, \ldots, m\}$.

While this modification avoids the theoretical issues with EVM, it brings its own. Where there are several categories, the distance between categories near opposite ends of the range becomes large, which in terms of deriving estimates from finite and particularly small data sets means that there are likely to be many zero or very small counts, even in both classes, since the probabilities of scoring either $ni(l-1) \wedge njk$ or $nil \wedge nj(k-1)$ are very small. These uninformative comparisons may lead to small or even zero results for individual estimators of the intervals between such pairs of individual thresholds compared to the real value. However, these uninformative estimators are treated the same as more informative ones, since the procedure makes no distinction when aggregating individual estimators into the final estimates. Considering this fact indicates that it would be desirable to account for uninformative estimators by excluding them or weighting them according to their informativeness.

Conditional Pairwise Adjacent Thresholds

An alternative approach to threshold estimation which avoids making comparisons between the extremes of the scale and permits the possibility of weighting estimators is to consider only pairs of adjacent thresholds. Setting $l = k + 1$ in Equation 10, we have:

$$\tau_{k+1} - \tau_k = \ln(\frac{P_{nik \wedge njk}}{P_{ni(k-1) \wedge nj(k+1)}}) + (\delta_i - \delta_j) \tag{14}$$

This provides an estimator for $\tau_{k+1} - \tau_k$ from any given pair of items (i, j), subject to eliminating the term $(\delta_i - \delta_j)$, which can be done in one of two ways. Firstly, following the logic employed in the modified EVM method, we can consider the item pair (j, i) rather than (i, j) and combine the two equations to give:

$$\tau_{k+1} - \tau_k = \frac{1}{2}[\ln(\frac{P_{nik \wedge njk}}{P_{ni(k-1) \wedge nj(k+1)}}) + \ln(\frac{P_{nik \wedge njk}}{P_{ni(k+1) \wedge nj(k-1)}})] \tag{15}$$

The second approach is simply to incorporate the item difficulty estimates, which can be estimated separately beforehand using PAIR or EVM, directly into Equation 14.

Once estimators have been derived from each pair of items, it remains to find a means of combining these estimators across all possible pairs of items in an appropriate way. The simplest approach to combining the estimators is to take the arithmetic mean, providing a final estimate for $\tau_{k+1} - \tau_k$. From Equation 15, we have:

$$\tau_{k+1} - \tau_k = \frac{1}{I(I-1)}\sum_{i<j}[\ln(\frac{P_{nik \wedge njk}}{P_{ni(k-1) \wedge nj(k+1)}}) + \ln(\frac{P_{nik \wedge njk}}{P_{ni(k+1) \wedge nj(k-1)}})] \tag{16}$$

Where I is the total number of items. Alternatively, from Equation 14, we have:

$$\tau_{k+1} - \tau_k = \frac{1}{I(I-1)}\sum_{i \neq j}[\ln(\frac{P_{nik \wedge njk}}{P_{ni(k-1) \wedge nj(k+1)}}) + (\delta_i - \delta_j)] \tag{17}$$

where I is the total number of items/criteria (we only need to consider cases where $i<j$ in Equation 15 since it is symmetrical in i and j). Where any of the counts are zero, neither Equation 17 nor Equation 16 will produce a finite estimate, so all such cases must be discarded (and the estimator count $I(I-1)$ adjusted accordingly); this means that cases where all item pairs produce a zero count, the method will fail to produce an estimate for $\tau_{k+1} - \tau_k$. This is more likely for Equation 16 since it contains three counts which may be zero, rather than two for Equation 17.

Weighting CPAT Estimators. Combining the set of estimators in the form of a simple arithmetic mean has certain disadvantages, as discussed for the modified formulation of EVM. Some estimators will be more stable than others since they contain more information; for instance, estimators drawn from pairs of items of similar difficulty are likely to have more observations than for estimators drawn from pairs of items of very different difficulties since persons are more likely to obtain similar scores on such pairs of items. However, noting that each estimator is a sufficient statistic in its own right, it is possible to take a weighted average of estimators rather than a simple arithmetic mean—in fact, any arbitrary set of weights will theoretically produce a valid estimate. We therefore have, from Equation 16:

$$\tau_{k+1} - \tau_k = \frac{\sum_{i<j} \omega_{ij}[\ln\frac{P_{nik \wedge njk}}{P_{ni(k-1) \wedge nj(k+1)}} + \ln\frac{P_{nik \wedge njk}}{P_{ni(k+1) \wedge nj(k-1)}}]}{\sum_{i<j} \omega_{ij}} \tag{18}$$

or, from Equation 17:

$$\tau_{k+1} - \tau_k = \frac{\sum_{i\neq j} \omega_{ij}[\ln \frac{P_{nik\wedge njk}}{P_{ni(k-1)\wedge nj(k+1)}} + (\delta_i - \delta_j)]}{\sum_{i\neq j} \omega_{ij}} \tag{19}$$

where ω_{ij} is the weight for item pair (i,j).

A standard approach to this weighting problem is to weight estimators by the inverse of their variance; this approach, known as inverse-variance weighting, draws its justification from the inequality $V[\theta] \geq 1/I[\theta]$ (Silvey, 1975): more information in an estimator corresponds to less variance. To calculate the inverse-variance weighting for an estimator, we must naturally calculate its variance. We note that the sufficient statistic to calculate the estimator is $\bar{X}_i = p$, where $X_i = \{x_{1i}, ..., x_{Ni}\}$, with $x_n = 1$ if person n scored (k, k) or 0 if person n scored $(k-1, k+1)$ among the N persons who scored either (k, k) or $(k-1, k+1)$ on the pair of items (i, j). We must calculate the variance of $g(\bar{X})$, where $g(Y) = ln(Y) - ln(1-Y)$. We can approximate this by using the delta method (Doob, 1935), which is based on Taylor series expansions:

$$V[g(Y)] \approx g'(E[Y])^2 V[Y] \tag{20}$$

Here, $Y = \bar{X}$, with $V[Y] = p(1-p)/N$ and $g'(Y) = 1/Y(1-Y)$, so we have:

$$V[g(Y)] \approx 1/Np(1-p) \tag{21}$$

which gives the weighting:

$$\omega_{ij} = Np(1-p) \tag{22}$$

There is a direct relationship here with the Fisher information function for the SLM, where $I = p(1-p)$. Indeed, this observation provides an alternative means of deriving the weighting function: note that $f^{-1}(y) = Ne^y/(1+e^y)$, which is identical to N dichotomous items of difficulty 0, the total Fisher information of which is $Np(1-p)$. The weights therefore represent weighting each estimator by the Fisher information of $f^{-1}(y)$.

If we now define n_{kk} as the count of $n_{ik} \wedge n_{jk}$ and $n_{(k-1)(k+1)}$ as the count of $n_{i(k-1)} \wedge n_{j(k+1)}$, we have $N = n_{kk} + n_{(k-1)(k+1)}$ and $p = n_{kk}/(n_{kk} + n_{(k-1)(k+1)})$. In terms of n_{kk} and $n_{(k-1)(k+1)}$, we have, from Equation 14:

$$\omega_{ij} = \frac{n_{kk} n_{(k-1)(k+1)}}{n_{kk} + n_{(k-1)(k+1)}} = \frac{1}{2} H(n_{kk}, n_{(k-1)(k+1)}) \tag{23}$$

where $H(n_{kk}, n_{(k-1)(k+1)})$ is the harmonic mean of n_{kk} and $n_{(k-1)(k+1)}$. Since we can multiply the top and bottom of Equations 18 and 19 by any arbitrary constant, we can simply set:

$$\omega_{ij} = H(n_{kk},\ n_{(k-1)(k+1)}) \tag{24}$$

Equation 18 contains a further complication, however: there are two terms, each with its own weighting, and these two weights must be further combined into a single weight. We should note that in doing this, we are combining two estimators which may be of different quality together via a straightforward arithmetic mean as per Equation 16, which cannot account for any difference in informativeness. There does not appear to be a clearly indicated way of combining weights, so following the spirit of Equation 24, the weighting is calculated by combining the terms into a single term and taking the harmonic mean of the numerator and denominator of the combined term as with Equation 24:

$$\ln\frac{P_{nik\wedge njk}}{P_{ni(k-1)\wedge nj(k+1)}}+\ln\frac{P_{nik\wedge njk}}{P_{ni(k+1)\wedge nj(k-1)}}=\ln\frac{{P_{nik\wedge njk}}^{2}}{P_{ni(k+1)\wedge nj(k-1)}P_{ni(k-1)\wedge nj(k+1)}} \tag{25}$$

Giving:

$$\omega_{ij}=H({n_{kk}}^{2},\ n_{(k-1)(k+1)}n_{(k+1)(k-1)}) \tag{26}$$

We name this Rasch–Andrich threshold estimation approach as the conditional pairwise adjacent thresholds method (CPAT); in order to distinguish between the two approaches to conditioning out the item difficulties from Equation 14, for the remainder of this paper we will call the first method (based on Equation 18) CPAT 1 and the second method (based on Equation 19) CPAT 2.

Additive Smoothing. There can be cases where estimation methods fail to produce a complete set of Rasch-Andrich thresholds since the sample is not large enough and/or poorly targeted across the range of item/threshold combinations, meaning that there is insufficient data to generate a full set of estimates—specifically, this is most likely to occur with estimates for extreme thresholds, that is threshold 1 and threshold m, where m is the maximum score. It is, however, possible to ensure that a set of estimates is returned for all samples by applying Choppin's (1985) alternative method for avoiding zeros in the PAIR conditional category frequency matrix, which is to add an arbitrary constant to all counts—this constant can be any small positive number, and does not need to be integer. This approach, called additive smoothing (Murphy, 2012), is designed to account for the fact that zero counts observed in a sample do not reflect underlying small but non-zero probabilities and amounts to the imposition of a weak (depending on the magnitude of the additive constant) uniform prior distribution. Although Choppin presents additive smoothing as an alternative to raising the matrix C to successive powers, the two approaches may be combined. An additive smoothing constant operates as a hyperparameter (Murphy, 2012), with its effect on estimates depending on the value chosen.

Additive smoothing has two effects on threshold estimates, which act in opposite directions. Since it imposes a uniform prior, it brings the quotient $P_{nik\wedge njk}/P_{ni(k-1)\wedge nj(k+1)}$ closer to one, which, after taking the logarithm, will reduce the magnitude of the threshold distance, which will tend toward reduced threshold distances (indeed, as the additive constant tends to infinity, all threshold distances will tend to 0). At the same time, additive smoothing leads to the inclusion of estimators with zero counts in the final estimates. These estimators, which must be excluded without additive smoothing, are the result of random error where one result is unlikely—for example, an estimator with 0 out of 20 cases of (0, 1) rather than 1 out of 20 where the true probability is .05. Excluding these estimators, and only these estimators, systematically results in bias toward reduced threshold distances since random error only in one direction is removed—estimators with random error in the other direction (2 or 3 out of 20 rather than 1 out of 20, which underestimate threshold distances, are included). In this way, additive smoothing removes a source of reduced threshold distances.

The net effect of these two effects is likely to vary with the value of the additive smoothing constant: any non-zero value will cause the second effect, although with information weighting, the effect will small for small additive smoothing constants. As the additive smoothing constant increases, the first effect will increase, suggesting that there may be an optimal value, which may vary across datasets.

Hypotheses for Testing

We hypothesize the following:

1. EVM will perform less well for threshold estimates than for overall item location estimates due to the issues outlined above.
2. CPAT will perform better than EVM.
3. Weighting CPAT estimators as described above will improve the performance of CPAT.
4. The use of an additive smoothing constant will improve performance in general up to optimal value, after which performance will worsen.

We do not hold active hypotheses regarding the relative performance of CPAT 1 and CPAT 2, or the relative performance of PAIR and EVM for item estimates; we leave these as exploratory questions to be answered. We will also compare the performance of CPAT to that of CMLE and JMLE.

Method

In order to evaluate the quality of the estimates against an objective baseline, it is necessary to use simulated rather than authentic data since it is only with simulated data that the original item parameters can be known. The recovered estimates can then be compared against the generating parameters using appropriate metrics, in which way direct comparisons between different estimation algorithms can be made. Luecht and Ackerman (2018) outline the broad approach to such simulation studies: select a model, specify item and person parameters, generate responses from parameters according to model, estimate parameters from the generated responses, and

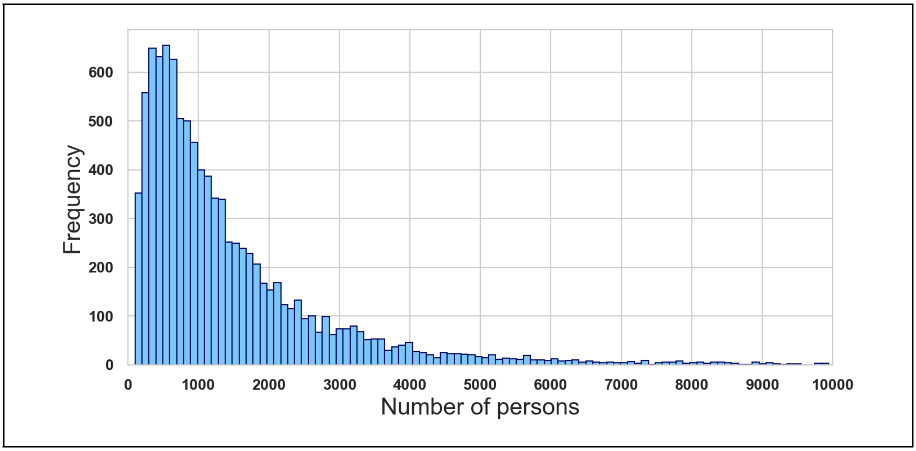

Figure 1. Sample size distribution.

compare the estimates to generating parameters. All analyses is conducted using code written by the authors in Python 3.

For this study, 10,000 different datasets were generated with varying numbers of persons and items and different generating parameters. The simulation parameters were generated following a seven-step procedure:

1. Generate a number of items from a discrete integer random uniform distribution in the range $[4,\ 10]$.
2. Generate a maximum score from a discrete integer uniform distribution in the range $[3,\ 7]$ (the number of Rasch–Andrich thresholds to be estimated).
3. Generate a set of item difficulties with a range drawn from a continuous uniform distribution in the range $[0.5,\ 3.5]$.
4. Generate a sample size in the range $[100,\ 10,\ 000]$ with a median of 1,000: from a discrete integer distribution of 10^x, where x is drawn from a truncated normal distribution in the interval [2, 4]. This generates a skewed distribution, the result of which is illustrated in Figure 1.
5. Generate a set of person abilities following a continuous normal distribution with a *SD* drawn from a continuous random uniform distribution in the range [1, 3.5].
6. Add an offset to the person ability distribution relative to the item difficulties from a continuous uniform distribution in the range [–0.5, 1].
7. Generate a set of thresholds with a mean category width from a continuous uniform distribution in the range [0.5, 2], with random perturbations from the mean. Only allow (but do not force) disordered thresholds in 50% of cases through the random perturbations, with a maximum disorder from a continuous uniform distribution in the range [0.5, 1].

The parameters were then used to generate simulated response sets following a five-step procedure:

1. Generate n person abilities, i items and m Rasch–Andrich thresholds as described above.
2. Calculate the category response probabilities P_{nik} for each person n on each item i for each threshold τ_k, where $k \in \{1, \ldots, m\}$.
3. Calculate the cumulative category response probabilities $P^{nik} = \sum_{j=0}^{k} P_{nij}$ for each person n on each item i for each threshold τ_k, where $k \in \{1, \ldots, m\}$.
4. Generate a table of numbers r_{ni} from a continuous uniform distribution in the range [0, 1].
5. Score the responses x_{ni} for each person n on each item i: for successive $k \in \{1, \ldots, m\}$, if r_{ni} is in the interval $(P^{ni(k-1)}, P^{nik}]$, set $x_{ni} = k - 1$.

The data were simulated to fit the RSM without data-model misfit in order to maximize the interpretability of comparisons between the generating parameters and the recovered estimates; the consequences of different kinds of item misfit therefore fall outside the scope of this study.

Two further datasets were generated from each dataset. Firstly, a reduced dataset was generated by removing 30% of persons at random. Secondly, a missing data dataset was generated by removing 30% of individual responses. Missing data can be classified according to the relationships between missingness and traits (Rubin, 1976); in the context of tests may be a design feature (computer adaptive tests, for example, always produce missing data). Here we only consider the simplest case of missing completely at random (MCAR).

The two additional datasets allow comparison between a reduction in the number of persons and the same reduction in individual responses, which provides an indication of a procedure's sensitivity to missing data—an algorithm which is sensitive to missing data will see a greater deterioration in accuracy under the missing data conditions, even though the datasets contain the same number of responses.

To compare the recovered estimates to generating parameters across a number of simulations, suitable summary statistics are required. Three summary statistics were calculated for this study to capture different aspects of the estimation error. Firstly, the root mean squared error (RMSE) of the point estimates provides a straightforward measure of the overall amount of error (RMSE is preferred here to MSE due to the interpretability of the units). Secondly, the *SD* ratio of the recovered item difficulty estimates to the original difficulty estimates used to generate the simulated data provides a measure of the relative dispersion of the two sets of estimates: $SD\,\text{ratio} = \sigma_1 / \sigma_2$, where σ_1 and σ_2 are the *SD*s of the two sets of estimates. *SD* ratio provides a means of investigating estimation bias—systemic error in the estimates. The same amount of estimation error with an *SD* ratio far from 1 is indicative of more estimation bias as opposed to random noise. Finally, to quantify the modeling error resulting from the combined item and threshold estimates, the parameter-estimation residuals (Luecht & Ackerman, 2018) are calculated. Parameter-estimation residuals are a means of separating the residual error due to parameter estimation error from that due to the stochastic noise which is an inherent feature of the model, and are calculated per response according to the following formula:

$$\varepsilon_2 = \hat{f} - f_1 = (x - f_1) - (x - \hat{f}) \tag{27}$$

where ε_2 is the parameter-estimation residual, x is the observed score, f_1 is the item response function (IRF) for the generating parameters and $\hat{f}$ is the IRF for the estimated parameters; in other words, ε_2 is the difference between the residual for the IRF of the generating parameters and the IRF of the estimated parameters. ε_2 depends on the targeting of the item relative to person ability; a very easy item where both f_1 and $\hat{f}$ are very high will result in a small value of ε_2 even when the difference between f_1 and $\hat{f}$ is relatively large, due the IRFs' asymptotic properties, whereas this will not be the case for items close to the person ability. In this way, ε_2 can be seen as a measure of how well the estimates model the specific dataset since unlike RMSE, which does not take into account the lack of information which can lead to large point estimate errors for items which are poorly targeted for the population, ε_2 implicitly accounts for this by quantifying the practical effect of estimation error. To provide a summary statistic for evaluation purposes, the RMSE of all ε_2 values for all responses in each simulation was calculated. For descriptive statistics, the primary reported summary statistic is the median rather than the mean. This is because the median is not sensitive to outliers, which are likely to be one-sided (RMSE and SD ratio both have a lower bound of 0 and SD ratio has a reciprocal relationship around 1 rather than a linear one)

To provide a statistical test for differences between distributions of summary statistics, the Wilcoxon (1945) signed-rank test, a non-parametric equivalent to the paired sample *t*-test which tests the null hypothesis that the medians of the two samples are the same and which does not require any assumptions of normality for the distributions (an assumption which is not met by the data in this study) was calculated. The Wilcoxon result is reported; as well as reporting *p*-values, which are likely to be trivially significant with such large sample sizes, meaningful interpretation is provided by effect sizes, which are not sensitive to sample size: firstly, the common language effect size (CLES; McGraw & Wong, 1992) is reported, which is equivalent to the area under the curve (AUC; Fawcett, 2003) and calculated according to the simple difference formula (Kerby, 2014). The CLES states the probability that, from a randomly chosen pair of samples, the result from the sample with the higher median will be higher than the result from the sample with the lower median. The standard Cohen's *d* effect size (Cohen, 1962) is not calculated directly since it assumes a normal distribution; instead, an equivalent Cohen's *d* value is derived from the CLES, following Ruscio's (2008) method of conversion. As a guide, Sawilowsky's (2009) expanded rules of thumb for interpretation of effect size are used: $d < 0.1$ is reported as trivial, $0.1 \leq d < 0.2$ as very small, $0.2 \leq d < 0.5$ as small, $0.5 \leq d < 0.8$ as medium, $0.8 \leq d < 1.2$ as large, $1.2 \leq d < 2$ as very large and $d < 2$ as huge. Results are compared when appropriate, both for the same method applied to different data conditions (full vs. reduced, full vs. missing and reduced vs. missing) and for different methods under the same data conditions (e.g., EVM vs. CPAT under full data conditions).

In additional to statistics, results are presented graphically, using scatter plots to compare results per simulation, with the identity line as a point of reference, and boxplots to compare distributions of results.

For the comparisons between CPAT and CMLE and JMLE, the same procedure is used, although on a smaller set of 100 simulations. The Winsteps program (Linacre, 2021) is used to generate CMLE and JMLE estimates. JMLE estimates include Wright's (1988) bias estimation correction.

Results

Overall Item Location Estimates

Table 1 summarizes the item difficulty estimation results for PAIR and EVM across the three data conditions. Figure 2 compares the performance of PAIR and EVM on each of the three data conditions; the Wilcoxon signed-rank test are as shown in Table 2.

The results suggest that least squares PAIR performs slightly better than EVM, although the absolute magnitude of the difference is small and EVM appears to be a viable alternative. All three effects are statistically significant, with $p<0.001$ in all three cases, although as previously discussed, this is perhaps a trivial observation for a study with so many datasets. The effect size is similar and small under full and reduced data conditions, and medium under the missing data condition, suggesting that EVM may be slightly more sensitive to missing data than PAIR.

Selecting PAIR as the better performing method, the Wilcoxon signed-rank test was conducted between the three possible pairs of data conditions, as shown in Table 3. Again $p<0.001$ in all three cases in Table 3. The effect size is huge in each case, indicating that although PAIR may be slightly less sensitive to missing data, there is still an effect—the distribution suggests around 10% more error with the same amount of data (30% removed) in the missing data pattern.

Figure 3 summarizes the performance of the least squares and EVM methods across the three data conditions, showing box-and-whisker plots featuring the median and 25th and 75th percentiles (boxes) and 2.5th and 97.5th percentiles (whiskers); outliers are not shown for clarity.

EVM Threshold Estimates

The results for the threshold estimates are shown in Table 4. The results are markedly less accurate than the results for the item difficulty estimates—the median RMSE for the threshold estimates under the full data condition is 0.439 logits, a figure which is relatively stable for the reduced data condition at 0.467 logits but rises

Table 1. Item Difficulty Estimation Results.

	Least squares		EVM	
Dataset	RMSE	*SD* ratio	RMSE	*SD* ratio
Full data	0.056	0.990	0.058	0.992
Reduced data	0.067	0.990	0.069	0.992
Missing data	0.074	0.993	0.077	0.996

Table 2. Wilcoxon Signed-Rank Test: Least Squares Versus EVM, Item Difficulties.

Dataset	Wilcoxon	*p*-Value	CLES	Cohen's *d*
Full data	31,138,136	<.001	0.623	0.443
Reduced data	31,278,833	<.001	0.626	0.453
Missing data	32,506,566	<.001	0.650	0.546

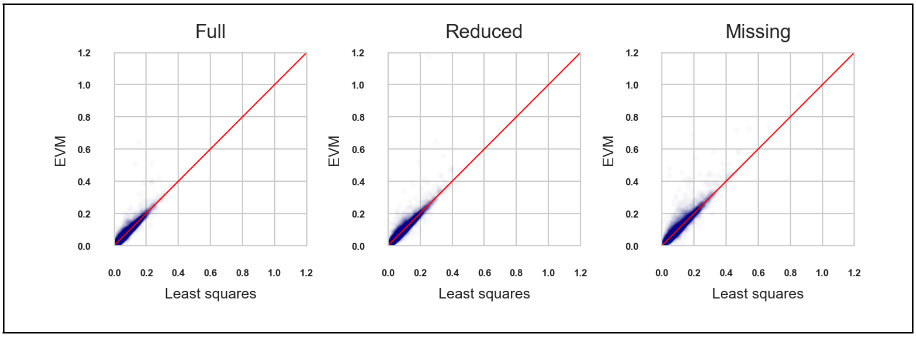

Figure 2. Comparison of RMSE of item estimates for PAIR versus EVM methods under full data, reduced data and missing data conditions.

as high as 1.166 logits for the missing data condition. A similar picture was observed for the *SD* ratio: the median is 1.278 for the full data condition; the figure remains stable for the reduced data condition at 1.292 but rises to 1.668 for the missing data condition. Figure 4, which plots the RMSE for each simulation under the three conditions pairwise, shows the deterioration in the quality of the estimates under missing data conditions graphically. Again, this is supported by the Wilcoxon signed-rank test, as shown in Table 5, with effect sizes well above 2 (huge) for comparisons between the missing data condition and both the full and reduced data condition.

Table 3. Wilcoxon Signed-Rank Test: Least Squares, Item Difficulties.

Dataset comparison	Wilcoxon	*p*-Value	CLES	Cohen's *d*
Full data vs. reduced data	41,177,061	<.001	0.824	1.314
Full data vs. missing data	44,355,857	<.001	0.887	1.714
Reduced data vs. missing data	32,622,770	<.001	0.653	0.555

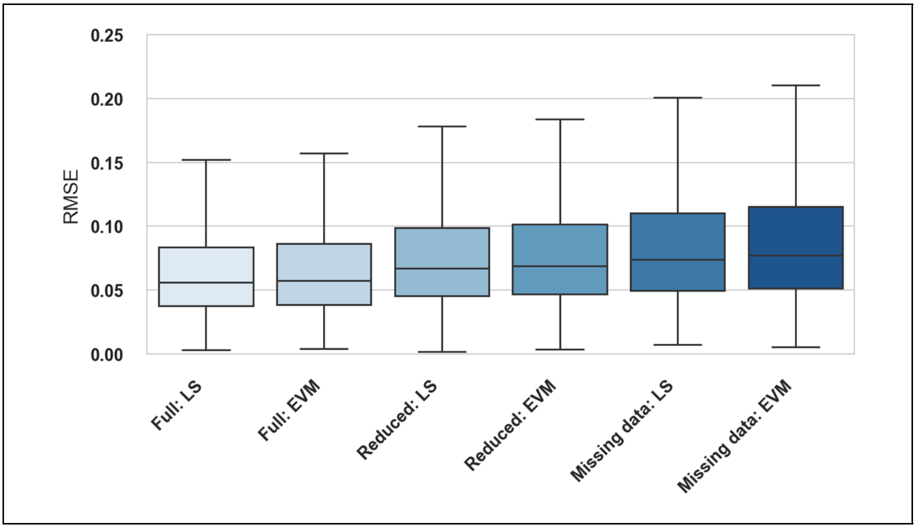

Figure 3. RMSE of item estimates for last squares (LS) and EVM under full data, reduced data, and missing data conditions.

CPAT Threshold Estimates

The results from the four CPAT variants on the 10,000 datasets under the three data conditions are shown in Table 6. It is immediately clear that the results in all four cases are far more accurate than EVM: in the full data case, the improvement in median RMSE is between 80.4% and 84.2% of the figure for EVM, falling slightly in the reduced data case to between 77.8% and 81.9%; in the missing data case they rise above 90% to between 90.0% and 91.9%. The results of the Wilcoxon tests for the worst performing CPAT variant (unweighted CPAT 1) versus EVM are shown in Table 7—effect sizes are all huge (between 4.08 and 6.18), and higher still for other CPAT variants.

In terms of comparisons between the four CPAT variants, Table 6 suggests that CPAT 2 performs better than CPAT 1 in both unweighted and weighted cases and that weighted methods perform better than unweighted methods for both CPAT 1

Table 4. EVM Rasch–Andrich Threshold Estimation Results.

Dataset	RMSE	*SD* ratio
Full data	0.439	1.278
Reduced data	0.467	1.292
Missing data	1.166	1.668

Table 5. Wilcoxon Signed-Rank Test: EVM, Rasch–Andrich Thresholds.

Dataset comparison	Wilcoxon	*p*-Value	CLES	Cohen's *d*
Full data vs. reduced data	28,774,124	<.001	0.772	1.056
Full data vs. missing data	37,217,155	<.001	0.999	4.406
Reduced data vs. missing data	37,133,334	<.001	0.997	3.860

Table 6. CPAT Rasch–Andrich Threshold Estimation Results.

	CPAT 1, unweighted		CPAT 1, weighted		CPAT 2, unweighted		CPAT 2, weighted	
Dataset	RMSE	*SD* ratio	RMSE	*SD* ratio	RMSE	*SD* ratio	RMSE	*SD* ratio
Full data	0.086	1.000	0.076	0.985	0.083	1.001	0.069	0.988
Reduced data	0.104	0.993	0.094	0.976	0.099	0.996	0.085	0.981
Missing data	0.117	0.982	0.110	0.964	0.108	0.988	0.095	0.975

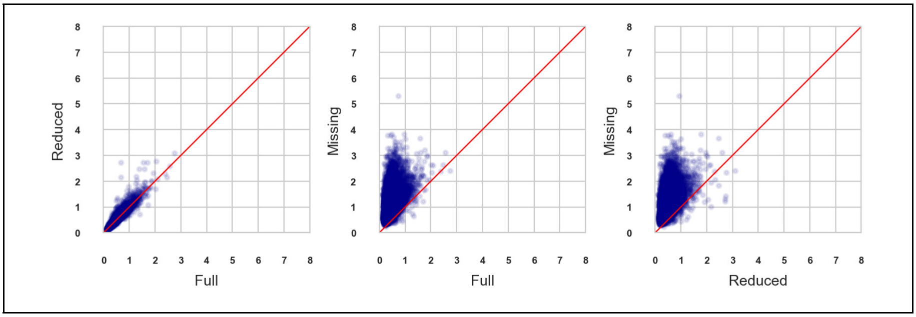

Figure 4. Comparison of RMSE of EVM Rasch–Andrich threshold estimates methods under full data, reduced data and missing data conditions.

and CPAT 2. This is confirmed by the results of the four relevant pairwise Wilcoxon tests, which are shown in Table 8. The best performing variant (weighted CPAT 2)

Table 7. Wilcoxon Signed-Rank Test: EVM Versus Unweighted CPAT 1 and, Rasch–Andrich Thresholds.

Dataset	Wilcoxon	*p*-Value	CLES	Cohen's *d*
Full data	37,227,572	<.001	0.999	4.555
Reduced data	37,178,813	<.001	0.998	4.082
Missing data	37,251,163	<.001	1.000	6.178

Table 8. Wilcoxon Signed-Rank Test: Rasch-Andrich Thresholds.

	Wilcoxon	*p*-Value	CLES	Cohen's *d*
Full				
CPAT 1 unwtd vs. CPAT 2 unwtd	25,975,853	<.001	0.697	0.731
CPAT 1 wtd vs. CPAT 2 wtd	23,714,007	<.001	0.637	0.494
CPAT 1 unwtd vs. CPAT 1 wtd	24,075,947	<.001	0.646	0.531
CPAT 2 unwtd vs. CPAT 2 wtd	25,281,707	<.001	0.679	0.656
CPAT 1 unwtd vs. CPAT 2 wtd	26,373,912	<.001	0.708	0.774
Reduced				
CPAT 1 unwtd vs. CPAT 2 unwtd	26,520,789.5	<.001	0.712	0.791
CPAT 1 wtd vs. CPAT 2 wtd	24,544,767	<.001	0.659	0.579
CPAT 1 unwtd vs. CPAT 1 wtd	22,970,900	<.001	0.617	0.420
CPAT 2 unwtd vs. CPAT 2 wtd	24,600,318	<.001	0.660	0.585
CPAT 1 unwtd vs. CPAT 2 wtd	26,313,886	<.001	0.706	0.768
Missing				
CPAT 1 unwtd vs. CPAT 2 unwtd	28,165,640.5	<.001	0.756	0.981
CPAT 1 wtd vs. CPAT 2 wtd	25,419,824	<.001	0.682	0.671
CPAT 1 unwtd vs. CPAT 1 wtd	21,961,703	<.001	0.590	0.320
CPAT 2 unwtd vs. CPAT 2 wtd	23,095,453	<.001	0.620	0.432
CPAT 1 unwtd vs. CPAT 2 wtd	26,418,621	<.001	0.709	0.779

Note. Unwtd = unweighted; wtd = weighted.

outperforms the worst performing variant (unweighted CPAT 1)with a large effect size across all three data conditions in both unweighted and weighted cases—in both cases, the effect size increases for the missing data case, suggesting that CPAT 2 is less sensitive to missing data than CPAT 1. Weighted variants outperform their unweighted variants with a small effect size across all three data conditions for both CPAT 1 and CPAT 2. The cumulative effect from unweighted CPAT 1 to weighted CPAT 2 results in medium (bordering on large), effect sizes of between 0.768 and 0.779 which remain stable across the data.

Figure 5 shows the relative performance of weighted CPAT 2—which, as the best performing variant we shall henceforth focus on and refer to simply as CPAT - versus

EVM on the three datasets, with the clearly superior performance of CPAT evident in all cases.

Figure 6 summarizes the performance of weighted CPAT against EVM across the three data conditions, showing box-and-whisker plots as for the item estimates in Figure 3. The 97.5th percentile for CPAT is considerably below the 25th percentile for EVM in all three cases.

Additive Smoothing

The percentages of cases where estimation failed to produce a complete set of Rasch–Andrich thresholds are described in Table 9. CPAT 2 has the lowest failure rate for all three data conditions. EVM has a lower failure rate than CPAT 1 for both the full and reduced data conditions (0.99% and 1.38% vs. 3.02% and 4.49%, but there is a marked deterioration in the performance of EVM for the missing data condition, with a failure rate of 11.13% compared to 6.3% for CPAT 1, further underlying EVM's sensitivity to missing data; CPAT 2 remains relatively unaffected by missing data, with the failure rate rising from 0.74% to 1.29%.

In order to investigate the consequences of applying this approach, the simulation study was repeated four times, with different additive smoothing constants: 0.01, 0.1, and 1. The distributions of the RMSE and *SD* ratios are shown in Figure 7 together with the original CPAT distribution (shown as an additive smoothing constant of 0), for those cases where CPAT returned estimates. The additive smoothing constant functions as a hyperparameter (Murphy, 2012) of the algorithm.

The results follow the hypothesized pattern—as well as ensuring that estimates are produced, additive smoothing appears to lead to better estimates up to a point, in particular by reducing estimation bias (*SD* ratio closer to 1). As Figure 7 shows, the median *SD* ratio approaches 1 as the additive smoothing constant increases, although the interquartile range and 95% ranges increase as the additive constant increases beyond 0.1, at which point RMSE also increases. A Wilcoxon test between the original CPAT estimates and those with the best-performing additive smoothing constant of $+0.1$ returns $p<0.001$ and $d=0.595$ (medium effect size).

Figure 8 shows the distributions of RMSE and *SD* ratio with a 0.1 additive constant, comparing the cases where CPAT failed to return an estimate without the additive constant with those where it was successful for the full data case. It is clear that the error of the estimates for the failed standard CPAT estimations is still large—this should perhaps be unsurprising given that these are all cases where the sample was poor for estimation purposes in size and/or targeting.

Parameter-Estimation Residuals

The RMS parameter-estimation residual for the combined item and threshold parameter set for EVM and CPAT, both with no additive smoothing and with an additive smoothing constant of 0.1, are shown in Table 10. For the full data and reduced data

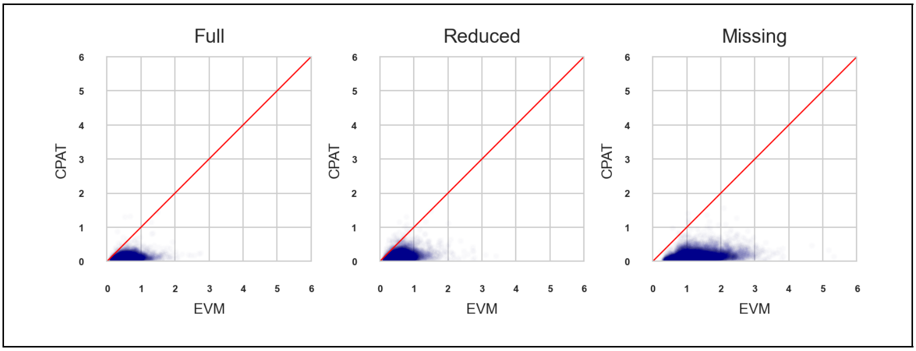

Figure 5. Comparisons of RMSE of Rasch–Andrich threshold estimates for EVM and CPAT under full data, reduced data, and missing data conditions.

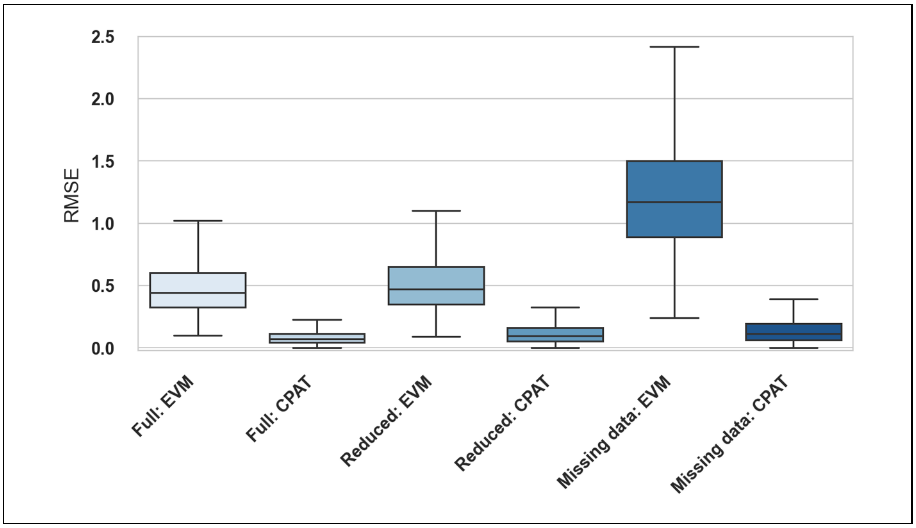

Figure 6. RMSE of Rasch–Andrich threshold estimates for EVM and CPAT under full data, reduced data, and missing data conditions.

Table 9. Estimation Failure Rate, Rasch–Andrich Thresholds.

	EVM (%)	CPAT 1 (%)	CPAT 2 (%)
Full	0.99	3.02	0.74
Reduced	1.38	4.49	0.91
Missing	11.13	6.3	1.29

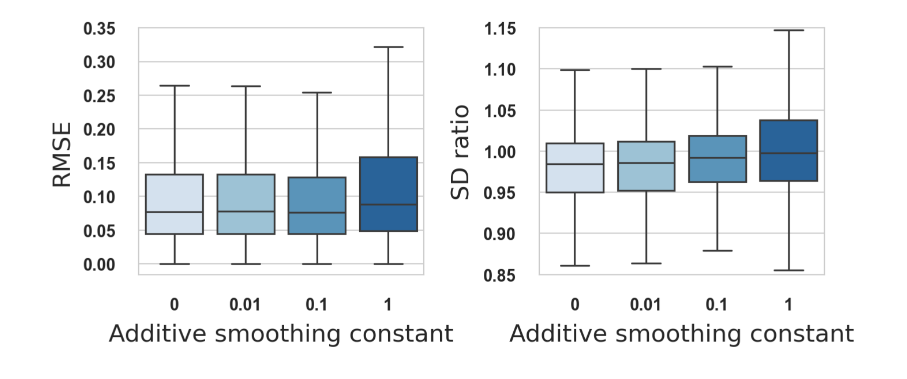

Figure 7. RMSE and *SD* ratio distributions for Rasch–Andrich threshold estimates using different additive smoothing constants.

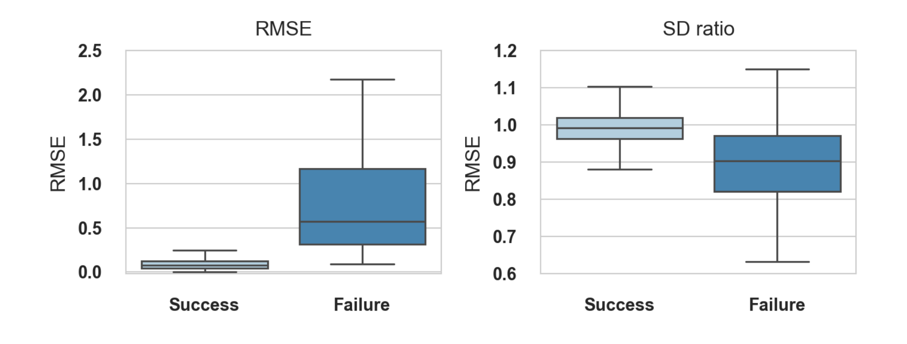

Figure 8. Comparison of distributions using additive smoothing (+0.1) between estimations which were successful and unsuccessful using CPAT with no additive smoothing.

conditions, CPAT performs almost identically with or without additive smoothing, and considerably better than EVM, which has a median RMS residual between 2.8 and 3.2 times larger. This disparity grows for the missing data condition between 5.7 and 6 times due to EVM's sensitivity to missing data; there is also a very marginal difference here in the performance of CPAT, although the median reduction in parameter-estimation residuals is only 0.25% for the full data condition, rising to 0.81% for the missing data condition.

Wilcoxon tests, shown in Table 11 support this, showing very large effect sizes between EVM and CPAT for full and reduced data conditions, and huge effect sizes for the missing data condition. The use of additive smoothing with CPAT causes a small effect for full and reduced data conditions and a medium effect size for the missing data condition, although it should be noted that practically the difference in the case of CPAT is almost trivial.

Table 10. Median Parameter-Estimation Residuals.

	Full	Reduced	Missing
EVM	0.165	0.177	0.417
CPAT, no AS	0.053	0.063	0.073
CPAT, AS	0.052	0.062	0.069

Note. AS = additive smoothing.

Table 11. Wilcoxon Signed-Rank Test: Parameter-Estimation Residuals and Method Comparisons.

	Wilcoxon	*p*-Value	CLES	Cohen's *d*
Full				
EVM vs. CPAT, no AS	35,554,603	<.001	0.916	1.950
EVM vs. CPAT, AS	35,681,605	<.001	0.919	1.981
CPAT, no AS vs. CPAT, AS	25,100,968	<.001	0.647	0.532
Reduced				
EVM vs. CPAT, no AS	35,332,685	<.001	0.910	1.899
EVM vs. CPAT, AS	35,523,590	<.001	0.915	1.943
CPAT, no AS vs. CPAT, AS	26,181,113	<.001	0.675	0.640
Missing				
EVM vs. CPAT, no AS	37,339,310	<.001	0.962	2.510
EVM vs. CPAT, AS	37,415,428	<.001	0.964	2.544
CPAT, no AS vs. CPAT, AS	27,268,753	<.001	0.703	0.752

Note. AS = additive smoothing.

Table 12 shows the results of Wilcoxon tests on the parameter-estimation residuals using the same method under different data conditions. The sensitivity of EVM to missing data is highlighted here; the effect size (Cohen's *d* equivalent) between the reduced data and missing data conditions, with the same total number of responses, is 2.617 (huge), which compares to 0.611 (no additive smoothing, medium) and 0.550 (additive smoothing medium). The effect sizes between full and reduced data conditions are fairy similar across all three methods, with EVM slightly lower (1.041) than CPAT (1.217 and 1.180); all three effect sizes are large.

CPAT versus CMLE and JMLE

Table 13 shows the median RMSE and *SD* ratio of the item and threshold estimates for CPAT (with a +0.1 additive smoothing constant), CMLE and JMLE (with estimation bias correction) under full, reduced, and missing data conditions. It is immediately clear that the results for both CPAT and CMLE are better than those for

Table 12. Wilcoxon Signed-Rank Test: Parameter-Estimation Residuals, Data Condition Comparisons.

	Wilcoxon	*p*-Value	CLES	Cohen's *d*
EVM				
Full vs. reduced	29,851,520.0	<.001	0.769	1.041
Full vs. missing	37,796,202.0	<.001	0.974	2.744
Reduced vs. missing	37,565,222.0	<.001	0.968	2.617
CPAT, no AS				
Full vs. reduced	31,252,939.0	<.001	0.805	1.217
Full vs. missing	33,864,186.0	<.001	0.873	1.610
Reduced vs. missing	25,894,124.0	<.001	0.667	0.611
CPAT, AS				
Full vs. reduced	30,969,200.0	<.001	0.798	1.180
Full vs. missing	33,421,124.0	<.001	0.861	1.535
Reduced vs. missing	25,279,800.0	<.001	0.651	0.550

Note. AS = additive smoothing.

JMLE across all datasets, with the difference being more pronounced for threshold estimation and under missing data conditions. The JMLE estimates for both items and thresholds generally display bias, in opposite directions: item estimates are stretched, while threshold estimates are compressed. Comparing CPAT with CMLE, we see that CMLE estimates are generally more accurate, although the differences are small—the largest difference between median RMSEs is for items under full data conditions and thresholds under missing data conditions, at 0.012 logits.

To consider the combined effect of item and threshold estimates, Table 14 shows the median parameter-estimation residuals for the three procedures under the three data conditions, with the distribution illustrated in boxplot form in Figure 9. We see the same distinction, with CPAT and CMLE performing similarly, particularly under reduced data conditions, with slightly better performance for CMLE, while the JMLE estimates are significantly less accurate. This is borne out by Wilcoxon tests, shown in Table 15: all tests between CPAT or CMLE and JMLE have a huge effect size, while those between CPAT and CMLE return medium to large effect sizes. All results are significant at $p<0.001$ except for CPAT versus CMLE on reduced data, where $p = 0.007$.

Discussion

Both PAIR and EVM showed good performance for the estimation of item difficulties. Their performance was similar enough to suggest that practically there is little difference, although the evidence suggests that the least-squares PAIR method marginally outperforms EVM. Both methods have some sensitivity to missing data, with results for the missing data condition containing significantly more error than under

Table 13. Median Estimation RMSE and *SD* Ratio: CPAT, CMLE, and JMLE.

	RMSE			*SD* ratio		
	CPAT	CMLE	JMLE	CPAT	CMLE	JMLE
Items						
Full	0.057	0.045	0.101	1.004	1.021	1.129
Reduced	0.062	0.053	0.106	1.003	1.022	1.128
Missing	0.063	0.066	0.160	1.012	1.033	1.232
Thresholds						
Full	0.075	0.071	0.324	1.015	0.979	0.850
Reduced	0.092	0.085	0.331	1.017	0.979	0.842
Missing	0.094	0.082	0.503	1.022	0.976	0.782

Table 14. Median Parameter-Estimation Residuals: CPAT, CMLE, and JMLE.

	CPAT	CMLE	JMLE
Full	0.048	0.036	0.110
Reduced	0.048	0.041	0.114
Missing	0.055	0.052	0.175

Table 15. Wilcoxon Signed-Rank Test: Parameter-Estimation Residuals: CPAT, CMLE, and JMLE.

	Wilcoxon	*p*-Value	CLES	Cohen's *d*
Full				
CPAT vs. CMLE	3,973.0	<.001	0.787	1.124
CPAT vs. JMLE	4,912.0	<.001	0.973	2.718
CMLE vs. JMLE	5,047.0	<.001	0.999	4.584
Reduced				
CPAT vs. CMLE	3,247.0	.007	0.643	0.518
CPAT vs. JMLE	4,978.0	<.001	0.986	3.097
CMLE vs. JMLE	5,048.0	<.001	1.000	4.745
Missing				
CPAT vs. CMLE	3,520.0	<.001	0.697	0.730
CPAT vs. JMLE	4,917.0	<.001	0.974	2.740
CMLE vs. JMLE	5,000.0	<.001	0.990	3.295

the reduced data condition, although this is relative, and from a low base (median RMSE rises from 0.056 logits to 0.074 for PAIR and 0.058 to 0.077 for EVM). This is to be expected from methods based on pairwise comparisons, since removing

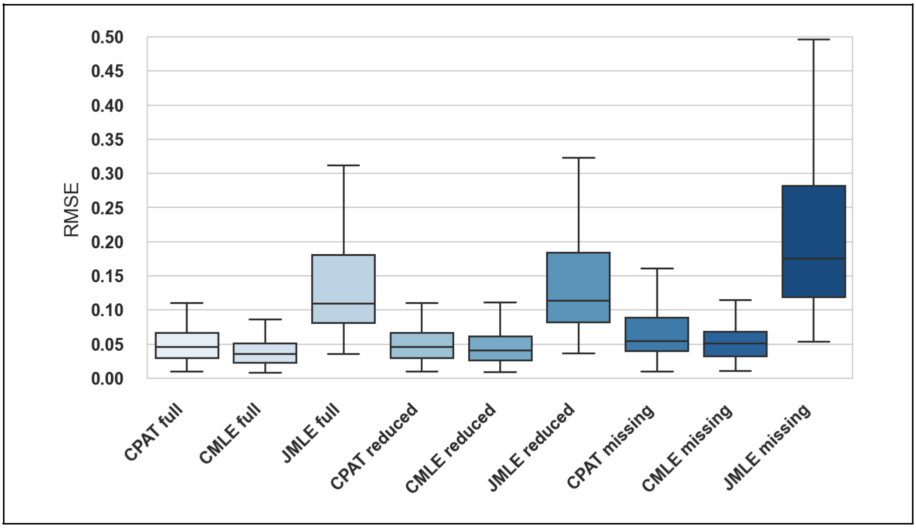

Figure 9. RMS parameter-estimation residuals: CPAT, CMLE, and JMLE.

individual data points from different response strings reduces the number of available comparisons by more than removing the same number of responses in whole strings: for example, removing a string of n responses removes $n(n-1)/2$ comparisons, while removing one response from each of n rows removes $n(n-1)$ comparisons—twice as many, with other missing patterns falling between these two extremes.

There are other methods of deriving the final difficulty vector, such as using a cosine maximization approach as described by Kou and Lin (2014), who also catalogue a variety of further methods developed for Saaty and Bennett (1977) and Saaty's (1994) AHP, or Choppin's (1985) iterative approach using higher matrix powers to add additional indirect comparisons until convergence. The evaluation of these further methods falls outside the scope of this study, but may be an avenue for further research both in terms of their performance and relative computational cost. It is worth noting in this regard that the least-squares approach is perhaps the simplest possible computationally as it requires only taking the arithmetic means of matrix rows.

The results showed support for hypothesis 1 in that the performance of EVM was markedly worse for threshold estimation than for item difficulty estimation, with a high degree of estimation bias, due to the formulation used to condition out person abilities and item difficulties simultaneously and the use of person raw scores to group persons, which introduces two issues: the use of person abilities in the estimation procedure introduces the incidental parameter problem, and additional error is introduced since the raw scores used for this are only proxies for person abilities, not the true underlying abilities. The performance of EVM was found to be particularly sensitive to missing data patterns.

The results showed support for hypothesis 2 in that CPAT in all its formulations significantly outperformed EVM. Compared to EVM, CPAT produced a median reduction in RMSE of between 84% and 92%, and a median reduction in parameter-estimation residuals of between 58% and 77%, with threshold estimation performance close to that for overall item locations. The CPAT formulation defined by Equation 16, where the overall item difficult estimates are eliminated entirely without inputting item estimates, appears to perform less well than that defined by Equation 17, where the overall item location estimates, with their own error of estimation, are used. The explanation for this likely resides in the issues around combining two estimators of different quality into a single estimator and how to weight these combined estimators, issues which do not arise in the case of CPAT 2.

The results showed support for hypothesis 3 in that weighting estimators produced an improvement in performance, with a median reduction of around 10% to 15% in RMSE. It is perhaps worthy of note that although weighting estimators results in a lower RMSE, the *SD* ratio is slightly farther from 1 in all cases; this suggestion of bias merits investigation. The weighting process interacts with random error distinctly in different directions—when random error leads to a reduced estimate of threshold distance, that is, overestimating the smaller count, this coincides with increased weighting; the converse is true for underestimates of smaller counts. In this way, there will typically be a slight bias in the estimates—specifically, threshold distances will be underestimated, although this particular source of bias should disappear asymptotically as sample size increases. In typical sample sizes, however, a small degree of systemic bias is introduced at the same time as reducing random error, although the trade-off appears to result in a net benefit.

The results also showed support for hypothesis 4 in that the use of additive smoothing produced a small improvement in performance, with a best median reduction of around 3% to 7% in RMSE for an additive smoothing constant of 0.1 (although the median reduction in parameter-estimation residuals was marginal, at less than 1%), as well as avoiding algorithm failure where data is sparse. Here it is worth noting that where no additive smoothing would have resulted in estimation failure, the estimates produced using additive smoothing were typically poor, which is perhaps unsurprising since this situation only occurred where the data were poor for estimation purposes. There may be scope for further gains from more sophisticated approaches to smoothing, although this falls outside the scope of this paper.

At this point, it may be worth reflecting beyond the immediate context on some general implications for pairwise comparison methods in general, and indeed any methods where observed counts are substituted directly for theoretical probabilities for estimation purposes. The superior performance of CPAT, despite using less data than EVM—only comparisons between adjacent thresholds are used, rather than comparisons across all pairs of thresholds—highlights the differences between theoretical probabilities, relating to infinite universes, and estimates of those probabilities drawn from finite sets of observations. In particular, where probabilities are small and zero observed counts can become a source of bias, even in larger datasets; the

CPAT approach, which involves discarding those comparisons where such counts are likely to occur, indicates that the quality of data can in some cases be more important than the quantity. The further use of additive smoothing can also, somewhat paradoxically, prevent compression of the set of estimates, despite its being the imposition of a uniform prior.

When comparing CPAT estimates with those obtained from CMLE and JMLE, CPAT comfortably outperformed JMLE in almost all cases, while CMLE estimates were more accurate than CPAT, although not by a large amount. There are, however, cases where this is not clear-cut, particularly as the amount of data is reduced; the item estimates under reduced data conditions were marginally more accurate than those from CMLE, and the CPAT parameter-estimation residuals under reduced data conditions were smaller than those for CMLE in 35.7% of cases. It is worth noting that the results do not support the hypothesis that CMLE estimates are sensitive to missing data (Eggen & Verhelst, 2006; Heine & Tarnai, 2015); rather, they suggest that it is JMLE estimates which display such sensitivity. Further research is required to determine whether the relative improvement in CPAT estimates on smaller datasets (reduced data conditions here) is indicative of a strong tendency for CPAT to perform better relative to CMLE as sample size decreases, or whether missingness is a meaningful factor on the relative performance of CPAT and CMLE.

The findings from this limited investigation suggest that the accuracy of CPAT is close to that of CMLE, and that there may be data designs where CPAT matches or even outperforms CMLE, while the trade-off in terms of computational speed versus a small performance penalty makes CPAT an attractive alternative in many contexts. We have not reported speed comparisons as part of this research since, although CPAT was considerably faster (the mean time for this study to load a dataset and compute PAIR item difficulty estimates and CPAT Rasch–Andrich threshold estimates was as little as 23.6 milliseconds), there are confounding factors which make direct numerical comparisons problematic—different software was used, written in different languages and run on different machines, with Winsteps also running additional analyses such as item fit as part of the estimation process.

Since all data for this study was generated to fit the model, further research is required to investigate the effect of misfitting data. Likewise, further research is required to investigate the effect of missing data which is not MCAR. The analytic framework presented in this study, with its evaluation of sensitivity to missing data (as opposed to reduced sample sizes) together with these further strands of enquiry, could constitute the basis of a comprehensive framework for the validation of estimation methods.

CPAT builds on the work of Choppin (1968, 1985) and Garner and Engelhard (2009) on non-iterative pairwise Rasch estimation procedures, combining the PAIR approach to overall item location estimation with a procedure which builds on the EVM approach to threshold estimation but does not simultaneously condition out person and item thresholds, instead using previously generated item estimates to create a chain of estimates for distances between adjacent threshold pairs. CPAT appears to constitute a highly computationally efficient approach to RSM parameter estimation that generates high quality estimates. One particular attraction of CPAT is its computational efficiency, since it is non-iterative and based on transformations of simple frequency counts. The results of this study provide evidence for CPAT as a fast estimation algorithm which approaches CMLE in its accuracy and represents an alternative to other more established procedures. CPAT would be particularly attractive in contexts where computational efficiency is at a premium, such as continuous item bank monitoring (e.g., tests at scale involving essays drawn from a large pool) and online systems with multiple simultaneous processes, and for the analysis of large datasets such as computer adaptive test banks. It also appears to have high applicability at the other end of the logistical scale where sample sizes may be small, such as preliminary validation studies of new essay questions and Likert-scale questionnaires, since the results of this study indicate that CPAT produces estimates of a comparable quality to those of CMLE, or sometimes even better, for such sample sizes. CPAT may also be attractive for contexts which involve missing data designs, since the estimates are not very sensitive to missing data, and these results suggest that they may be considerably less sensitive than JMLE estimates in particular. As such, there are a wide range of practical contexts where CPAT could be seen as a viable alternative to other established estimation procedures.

Acknowledgments

The authors would like to thank Edmund Jones, Andrew Caines and two anonymous reviewers whose insightful comments helped improve this manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This paper reports on research supported by Cambridge Assessment, University of Cambridge.

ORCID iD

Mark Elliott https://orcid.org/0000-0003-3302-5477

References

Andersen, E. (1970). Asymptotic Properties of Conditional Maximum-Likelihood Estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, *32*(2), 283–301.

Andersen, E. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26*, 31–44.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Andrich, D., & Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement*, *4*(3), 205–221.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

Choppin, B. (1968). Item bank using sample-free calibration. *Nature*, *219*(5156), 870–872.

Choppin, B. (1985). A fully conditional estimation procedure for Rasch model parameters. *Evaluation in Education*, *9*(1), 29–42.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145–153.

Doob, J. L. (1935). The Limiting Distributions of Certain Statistics. *Annals of Mathematical Statistics*, *6*(3), 160–169.

Eggen, T. J., & Verhelst, N. D. (2006). Loss of information in estimating item parameters in incomplete designs. *Psychometrika*, *71*(2), 303–322.

Fawcett, T. (2003). *ROC graphs: Notes and practical considerations for data mining researchers* Technical Report HPL-2003–4, Palo Alto, CA: HP Laboratories.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *46*(1), 59–77.

Garner, M., Jr. (2002). An eigenvector method for estimating item parameters of the dichotomous and polytomous Rasch models. *Journal of Applied Measurement*, *3*(2), 107–128.

Garner, M., & Engelhard, G., Jr. (2009). Using paired comparison matrices to estimate parameters of the partial credit Rasch measurement model for rater-mediated assessments. *Journal of Applied Measurement*, *10*(1), 30–41.

Ghosh, M. (1994). Inconsistent maximum likelihood estimators for the Rasch model. *Statistics & Probability Letters*, *23*(2), 165–170.

Heine, J. H., & Tarnai, C. (2015). Pairwise Rasch model item parameter recovery under sparse data conditions. *Psychological Test and Assessment Modeling*, *57*(1), 3–36.

Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, *3*(1), 11.IT.*3*.1.

Kou, G., & Lin, C. (2014). A cosine maximization method for the priority vector derivation in AHP. *European Journal of Operational Research*, *235*(1), 225–232.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, *95*(2), 391–413.

Linacre, J. M. (1994). Many-facet Rasch measurement. Chicago: MESA Press.

Linacre, J. M. (2021). *Winsteps* (Version 5.1.1) [Computer software]. Winsteps.com. https://www.winsteps.com/

Luecht, R., & Ackerman, T. A. (2018). A technical note on IRT simulation studies: Dealing with truth, estimates, observed data, and residuals. *Educational Measurement Issues and Practice*, *37*(3), 65–76.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361–365.

Murphy, K. P. (2012). *machine learning: A probabilistic perspective*. The MIT Press.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*(1), 1–32.

Rasch, G. (1960). *Probabilistic* m*odels for* s*ome intelligence and attainment* t*ests*. Danmarks Paedogogiske Institut.

Rasch, G. (1961). On general laws and meaning of measurement in psychology [Symposium]. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 321–333). University of California Press.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581–592.

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30.

Saaty, T. L. (1994). How to make a decision: The analytic hierarchy process. *Interfaces*, *24*(6), 19–43.

Saaty, T. L., & Bennett, J. P. (1977). A theory of analytical hierarchies applied to political candidacy. *Behavioral Sciences*, *22*(4), 237–245.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 597–599.

Silvey, S. D. (1975). *Statistical inference*. Chapman and Hall.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

Wright, B. (1988). The efficacy of unconditional maximum likelihood bias correction: Comment on Jansen, van den Wollenberg, and Wierda. *Applied Psychological Measurement*, *12*(3), 315–318.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis Rasch measurement*. MESA Press.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*(1), 23–48.

Zermelo, E. (1929). Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, *29*(1), 436–460.

Zwinderman, A. H. & van den Wollenberg, A. L. (1990). Robustness of Marginal Maximum Likelihood Estimation in the Rasch Model. *Applied Psychological Measurement*, *14*(1), 73–81.