

A Note on Tutz's Pairwise Separation Estimator

Alexander Robitzsch ^{1,2} 

¹ IPN—Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany; robitzsch@leibniz-ipn.de

² Centre for International Student Assessment (ZIB), Olshausenstraße 62, 24118 Kiel, Germany

Abstract

The Rasch model has the desirable property that item parameter estimation can be separated from person parameter estimation. This implies that no assumptions about the ability distribution are required when estimating item difficulties. Pairwise estimation approaches in the Rasch model exploit this principle by estimating item difficulties solely from sample proportions of respondents who answer item i correctly and item j incorrectly. A recent contribution by Tutz introduced Tutz's pairwise separation estimator (TPSE) for the more general class of homogeneous monotone (HM) models, extending the idea of pairwise estimation to this broader setting. The present article examines the asymptotic behavior of the TPSE within the Rasch model as a special case of the HM framework. It should be emphasized that both analytical derivations and a numerical illustration show that the TPSE yields asymptotically biased item parameter estimates, rendering the estimator inconsistent, even for a large number of items. Consequently, the TPSE cannot be recommended for empirical applications.

Keywords: item response model; parameter separability; Rasch model; monotone homogeneity model; Tutz's pairwise separation estimator

1. Introduction

In the social sciences, cognitive tests or questionnaires are often administered that yield $I > 1$ binary (i.e., dichotomous) random variables X_i for $i = 1, \dots, I$. These variables are frequently analyzed using item response theory (IRT; [1–5]) models, which reduce the dimensionality of the item responses to a unidimensional or multidimensional latent variable θ . Unidimensional IRT models [6] are particularly appealing because they summarize the I item responses X_i into a single latent variable θ .

The earliest development in IRT was the Rasch model [7–11], which models the dichotomous item response of person n on item i as

$$p_{ni} = P(X_{ni} = 1) = \Psi(\theta_n - b_i) = \frac{\exp(\theta_n - b_i)}{1 + \exp(\theta_n - b_i)} = \Psi(\theta_n - b_i), \quad (1)$$

where θ_n denotes the ability of person n and b_i the difficulty of item i . Note that the item response function (IRF) in the Rasch model employs the logistic link function Ψ .

The IRF in (1), which relies on the logistic link function Ψ , can be replaced by any smooth monotone link function F , yielding item response probabilities of the form

$$p_{ni} = P(X_{ni} = 1) = F(\theta_n - b_i). \quad (2)$$

The IRT model in (2) is also referred to as a homogeneous monotone (HM; [5,12,13]) model.



Received: 22 November 2025

Revised: 31 December 2025

Accepted: 5 January 2026

Published: 13 January 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

The Rasch model defined in (1) has the advantage, compared with the IRT model in (2), that the sum score $\sum_{i=1}^I X_{ni}$ is a sufficient statistic for θ_n when the Rasch model is estimated by maximum likelihood [14]. In addition, estimators of item difficulties can be constructed without simultaneously estimating the ability parameter θ_n . Unfortunately, this property is often conflated with concepts of parameter separability or specific objectivity in the Rasch model [15,16]. Conditional maximum likelihood estimation [9,14] and pairwise estimation approaches [17–21] both build on this idea.

To illustrate the conditioning principle, consider the ratio

$$\frac{P(X_{ni} = 1, X_{nj} = 0)}{P(X_{ni} = 1, X_{nj} = 0) + P(X_{ni} = 0, X_{nj} = 1)} = \frac{\exp(-b_i)}{\exp(-b_i) + \exp(-b_j)}. \quad (3)$$

It is evident from (3) that the ability parameter θ_n cancels out, implying that item parameters b_i can be estimated independently of person parameters in an estimation framework based on such ratios. Furthermore, it follows (see, e.g., [22]) that

$$\frac{P(X_{ni} = 1, X_{nj} = 0)}{P(X_{ni} = 0, X_{nj} = 1)} = \frac{\exp(-b_i)}{\exp(-b_j)} = \exp(-(b_i - b_j)). \quad (4)$$

The identities in (3) and (4) allow for the estimation of item difficulties b_i from probabilities $f_{ij} = P(X_i = 1, X_j = 0)$ for item pairs (i, j) in samples of persons. The row averaging estimation approach of Choppin [17] relies on the pairwise estimation principle and uses log-transformed ratios of sample proportions (see (4)) as

$$\log \frac{f_{ij}}{f_{ji}} = -(b_i - b_j). \quad (5)$$

Hence, estimates of the probabilities f_{ij} can be used to estimate item difficulties.

In a recent paper, Tutz [23] (see also [5,24]) proposed Tutz's pairwise separation estimator (TPSE) for estimating item difficulties in the more general HM model defined in (2). The TPSE is motivated by the idea that person and item parameter estimation can be separated. Following this reasoning, the estimator uses sample proportion estimates of f_{ij} and constructs item difficulty estimates based on pairwise differences. However, Tutz [5,23] did not provide conditions under which the TPSE yields consistent item parameter estimates, that is, the asymptotic behavior was not examined. The original reference [23] included a simulation study, but it relied on small samples and limited variation in item and person parameter distributions, as well as a small number of items. In particular, the simulation study did not offer evidence regarding the consistency of the TPSE.

The present article investigates the statistical properties of the TPSE in greater depth. The focus lies on the asymptotic case, that is, on determining the bias of the TPSE in infinite sample sizes. Attention is restricted to the Rasch model, because this model permits more tractable derivations of the bias, and the expressions involved in the TPSE simplify under this well-studied HM model compared to the general HM specification with an arbitrary smooth monotone link function F . It is shown in the present article that the TPSE generally produces biased item parameter estimates and therefore cannot be recommended as a reliable estimator in the Rasch model. As a consequence, item parameter estimates in the HM model are also biased. Although the formulation of the TPSE based on the parameter separability principle in [23] may have some aesthetic appeal, it is not useful as a reliable estimation technique.

The remainder of the article is organized as follows. Section 2 reviews the TPSE as introduced in Tutz [23]. Section 3 examines the bias of the TPSE item difficulty estimates in the Rasch model using analytical derivations. A Numerical Illustration is presented in Section 4 to confirm the analytical findings. An empirical example is additionally provided in Section 5. Finally, Section 6 provides a concluding discussion.

2. Tutz's Pairwise Separation Estimator

In this section, the TPSE as proposed in [23] is presented. Assume that the HM IRT model with link function F holds. If the item response probabilities $p_{ni} = P(X_{ni} = 1)$ were available, one could compute the difference

$$F^{-1}(p_{ni}) - F^{-1}(p_{nj}) = -(b_i - b_j). \quad (6)$$

By conditioning on the same person n in (6), the ability parameter θ_n is removed. Hence, the estimation of b_i is separated from the estimation of θ_n .

However, only binary observations $X_{ni} \in \{0, 1\}$ are available instead of the probabilities p_{ni} . Consequently, the difference in (6) cannot be applied directly to the binary data because the ability values θ_n are unknown. To address this issue, ref. [23] introduced pseudo-observations X_{ni}^* defined as

$$X_{ni}^* = \gamma(1 - X_{ni}) + (1 - \gamma)X_{ni}. \quad (7)$$

An incorrect item response $X_{ni} = 0$ is replaced by $\gamma \in (0, 0.5)$ in the pseudo-observation X_{ni}^* , while a correct response $X_{ni} = 1$ becomes $1 - \gamma \in (0.5, 1)$. The unknown probabilities p_{ni} in (6) are thus replaced by the pseudo-observations X_{ni}^* , leading to

$$\hat{b}_j(n) - \hat{b}_i(n) := F^{-1}(X_{ni}^*) - F^{-1}(X_{nj}^*) = \begin{cases} 0 & \text{if } X_{ni} = X_{nj} \\ \gamma_1 - \gamma_0 & \text{if } X_{ni} = 1, X_{nj} = 0 \\ -(\gamma_1 - \gamma_0) & \text{if } X_{ni} = 0, X_{nj} = 1 \end{cases}, \quad (8)$$

where $\gamma_1 = F^{-1}(1 - \gamma)$ and $\gamma_0 = F^{-1}(\gamma)$. According to Tutz [23], the right-hand side of (8) serves as an empirical approximation of $F^{-1}(p_{ni}) - F^{-1}(p_{nj})$. The parameter difference $\hat{b}_j(n) - \hat{b}_i(n)$ in (8) now has an explicit notation that indicates its dependence on person n .

By defining $\gamma_{10} = \gamma_1 - \gamma_0$, it follows from (8)

$$\hat{b}_j(n) - \hat{b}_i(n) = \gamma_{10} \frac{P(X_{ni} = 1, X_{nj} = 0) - P(X_{ni} = 0, X_{nj} = 1)}{P(X_{ni} + X_{nj} = 1)}. \quad (9)$$

Thus, according to Tutz [23], the difference $\hat{b}_j(n) - \hat{b}_i(n)$ is intended to estimate $\delta_{ij} = b_j - b_i$ by aggregating contributions across persons n :

$$\hat{\delta}_{ji} = \hat{b}_j - \hat{b}_i = \gamma_{10} \frac{\hat{f}_{ij} - \hat{f}_{ji}}{\hat{f}_{ij} + \hat{f}_{ji}}, \quad (10)$$

where $\hat{f}_{ij}$ is the proportion of persons who responded correctly to item i and incorrectly to item j and thus estimates population probabilities $f_{ij} = P(X_{ni} = 1, X_{nj} = 0)$. By the law of large numbers, it follows that $\hat{f}_{ij}$ are consistent estimators of f_{ij} . For symmetric link functions F , it holds that $\gamma_{10} = 2\gamma_1 = 2F^{-1}(1 - \gamma)$.

The TPSE estimate of b_i is finally defined in [23] as

$$\hat{b}_i = \frac{1}{I} \sum_{j=1}^I (\delta_{ij} - \delta_{1j}). \quad (11)$$

This construction yields $\hat{b}_1 = 0$, meaning that the first item serves as the reference item with fixed difficulty 0.

It should be emphasized that the pseudo-observations are introduced by Tutz [23] solely to derive the TPSE. The pseudo-observations are not explicitly used in the estimation procedure; instead, only a relationship between $\hat{\delta}_{ji}$ and the sample proportions $\hat{f}_{ij}$ is derived.

The dependence of the item parameter estimates obtained from the TPSE in (11) on the nuisance parameter γ is explicit. Tutz proposes several data-driven criteria for determining optimal values of the nuisance parameter γ . Note that an optimal γ value must be estimated in an additional step to obtain item difficulty estimates with the most favorable properties [23].

The following Section 3 investigates the bias of the item difficulty estimate $\hat{b}_i$ produced by the TPSE as a function of γ and discusses the optimal choice of this parameter to minimize bias in the estimated item difficulties. The focus is not on determining optimal γ values in concrete samples; instead, an optimistic best-case choice of γ that minimizes the average absolute bias in item parameter estimates is considered. It is shown that even in this best-case scenario, the TPSE yields substantially biased item parameter estimates, rendering the estimator rarely suitable for empirical research.

3. Bias Derivation of the Tutz's Pairwise Separation Estimator (TPSE) in the Rasch Model

In this section, the bias in the estimated item difficulties $\hat{b}_i$ obtained from the TPSE is examined. The asymptotic properties of the TPSE were not thoroughly investigated in [23]. Furthermore, the original source did not provide a convincing simulation study demonstrating satisfactory asymptotic behavior of the proposed estimator. Instead, the TPSE was introduced in a largely heuristic manner.

This section derives the bias in the case of the logistic link function $F = \Psi$, which corresponds to the Rasch model. If satisfactory statistical properties cannot be obtained in this setting, similar issues can be expected for other link functions in HM models. In the following derivation, the bias in item parameter estimates is assumed to be derived under the condition that θ follows an unknown distribution under random sampling [25]; that is, θ is integrated out. The item difficulty estimate in (11) depends critically on the quantities $\hat{\delta}_{ji}$ for item pairs (i, j) . These quantities, in turn, rely on the sample proportions $\hat{f}_{ij}$ (see (10)), representing the proportion of persons who answered item i correctly and item j incorrectly.

Assume that the sample proportions $\hat{f}_{ij}$ converge to the corresponding population probabilities f_{ij} . Further assume that these probabilities are obtained in the Rasch model for persons with identical ability $\theta = \theta_n$. Then, the corresponding population quantity δ_{ji} can be expressed as

$$\delta_{ji} = \frac{f_{ij} - f_{ji}}{f_{ij} + f_{ji}} = \frac{\exp(-b_i) - \exp(-b_j)}{\exp(-b_i) + \exp(-b_j)} = \frac{1 - \exp(-(b_j - b_i))}{1 + \exp(-(b_j - b_i))} = H(b_j - b_i), \quad (12)$$

where the function H is introduced for convenience. Note that $H(x) = \tanh(x/2)$.

Using the notation $\gamma_{10} = 2\Psi^{-1}(\gamma) = 2\gamma_1$, the TPSE expression in (10) becomes, for the Rasch model,

$$\delta_{ji} = 2\gamma_1 H(b_j - b_i). \quad (13)$$

The quantity δ_{ji} is intended to estimate the difference $b_j - b_i$. Hence, it is essential to evaluate the right-hand side of (13) and determine how it deviates from the target quantity $b_j - b_i$. Importantly, there cannot exist a unique value of γ_1 such that the right-hand side of (13) equals $b_j - b_i$. Otherwise, there would exist a value of γ_1 for which $2\gamma_1 \tanh(x/2) = x$ for all real values of x , which is impossible because $\tanh$ does not have a constant derivative.

To gain further insight into the emergence of bias in the item parameter estimates of the TPSE, a Taylor expansion of H around $x = 0$ up to third order based on the $\tanh$ function is employed. Terms up to third order are used to simplify the presentation and to highlight the main contributors to the bias when item difficulty differences do not deviate too strongly from 0. The Taylor approximation of H is given by

$$H(x) \simeq \frac{1}{2}x - \frac{1}{24}x^3. \quad (14)$$

To quantify the bias in $\hat{\delta}_{ji}$, the biasing term in the item-pair estimate for pair (i, j) is defined as

$$B(x) = 2\gamma_1 H(x) - x. \quad (15)$$

Using the Taylor expansion in (14) gives

$$B(x) \simeq (\gamma_1 - 1)x - \frac{\gamma_1}{12}x^3. \quad (16)$$

The expression in (16) can now be used to derive the bias in the estimated item difficulties $\hat{b}_i$ of the TPSE. From (11),

$$\hat{b}_i = \frac{1}{I} \sum_{j=1}^I (\hat{\delta}_{ij} - \hat{\delta}_{1j}) = b_i - b_1 + \frac{1}{I} \sum_{j=1}^I [B(b_i - b_j) - B(b_1 - b_j)]. \quad (17)$$

Hence, the bias simplifies to

$$\text{Bias}(\hat{b}_i) = \frac{1}{I} \sum_{j=1}^I [B(b_i - b_j) - B(b_1 - b_j)] \simeq (\gamma_1 - 1)b_i - \frac{\gamma_1}{12I} \sum_{j=1}^I [(b_i - b_j)^3 + b_j^3], \quad (18)$$

which further reduces, using $b_1 = 0$, to

$$\text{Bias}(\hat{b}_i) \simeq \left(\gamma_1 - 1 - \frac{\gamma_1}{4}s_2 \right) b_i + \frac{\gamma_1 s_1}{4} b_i^2 - \frac{\gamma_1}{12} b_i^3, \quad (19)$$

where $s_k = I^{-1} \sum_{j=1}^I b_j^k$ for $k = 1, 2$ denote the first two moments of the distribution of true item difficulties b_i .

The bias Formula (19) highlights the factors that influence the bias. Because b_i^2 is larger than 0 for $b_i \neq 0$, this component of the bias occurs whenever s_1 differs from 0; that is, when the average item difficulty is not equal to 0. This contribution can be reduced by selecting a reference item with $b_1 = 0$ such that s_1 becomes 0. This is likely achieved by choosing an item with medium item difficulty. Under the condition $s_1 = 0$, γ_1 must be chosen to balance the effects of b_i and b_i^3 in (19). However, it becomes clear that no universal choice of γ_1 can produce a bias of 0 for $\hat{b}_i$ for all possible values of b_i .

Assuming $s_1 = 0$ in (19) gives

$$\text{Bias}(\hat{b}_i) = \left(\gamma_1 - 1 - \frac{\gamma_1}{4}s_2 \right) b_i - \frac{\gamma_1}{12} b_i^3. \quad (20)$$

A possible strategy for selecting γ_1 in a given test is to minimize

$$\mathcal{B}(\gamma_1) = E_i[|\text{Bias}(\hat{b}_i)|], \quad (21)$$

where the expectation is interpreted as averaging across a (hypothetical) distribution of item parameters.

If all item difficulties b_i are substantially smaller than 1 in absolute value, the linear term in (20) dominates. In this case,

$$\gamma_1 = \Psi^{-1}(1 - \gamma) = \frac{1}{1 - s_2/4}. \quad (22)$$

3.1. Bias of the TPSE Under a Uniform Distribution of Item Difficulties

To illustrate (22), assume that the true item difficulties b_i follow a uniform distribution on $[-L, L]$. Then $s_1 = E_i(b_i) = 0$ and $s_2 = E_i(b_i^2) = L^2/3$, which yields $\gamma_1 = 1/(1 - L^2/12)$. For $L = 0.5$, this gives $\gamma_1 = 1.021$ and $\gamma = \Psi(1 - \gamma_1) = 0.264$. For $L = 1.0$, the values are $\gamma_1 = 1.091$ and $\gamma = 0.252$. For $L = 1.5$, the values are $\gamma_1 = 1.231$ and $\gamma = 0.226$. For $L = 2.0$, the result is $\gamma_1 = 1.500$ and $\gamma = 0.182$.

Now consider the more general case in which the term b_i^3 in (20) cannot be neglected, while still assuming that $s_1 = 0$. The bias in the estimated difficulties is again illustrated under the assumption that item difficulties follow a uniform distribution on $[-L, L]$. Further assume that $L < \sqrt{6} = 2.449$. Define $k = 1 - \frac{L^2}{12}$ and consider the range $\frac{1}{k} < \gamma_1 < \frac{1}{1-L^2/6}$. Under these conditions, the average absolute bias takes the form

$$\mathcal{B}(\gamma_1) = -\frac{\gamma_1 k L}{2} + \frac{L}{2} + \frac{\gamma_1 L^3}{48} + \frac{6}{\gamma_1 L} (\gamma_1 k - 1)^2 L \quad (23)$$

The optimal value of γ_1 is given by

$$\gamma_1^* = \frac{1}{k} \sqrt{\frac{1}{1 - \frac{L^2(8-L^2)}{96k^2}}} = \frac{1}{k} \sqrt{\frac{1}{1-D}} \quad (24)$$

with $D = \frac{L^2(8-L^2)}{96k^2}$. The minimum value of the average absolute bias is

$$\mathcal{B}^* = \frac{L}{2} - \frac{L}{2\sqrt{D}} + \frac{L^3}{48k\sqrt{D}} + \frac{6k}{L\sqrt{D}} (1 - \sqrt{D})^2 \quad (25)$$

Evaluating (25) yields the following numerical values. For $L = 0.5$, the results are $\gamma_1^* = 1.032$, $\gamma^* = 0.263$, and $\mathcal{B}^* = 0.001$. For $L = 1.0$, the values are $\gamma_1^* = 1.142$, $\gamma^* = 0.242$, and $\mathcal{B}^* = 0.012$. For $L = 1.5$, the results are $\gamma_1^* = 1.380$, $\gamma^* = 0.201$, and $\mathcal{B}^* = 0.049$. For $L = 2.0$, the values are $\gamma_1^* = 1.897$, $\gamma^* = 0.130$, and $\mathcal{B}^* = 0.162$.

To validate the analytical derivations, $\mathcal{B}(\gamma_1)$ was also evaluated numerically to identify optimal values. Figure 1 presents the corresponding average absolute bias as a function of γ_1 , and the numerical optimization of $\mathcal{B}$ with respect to γ_1 coincided with the analytical results.

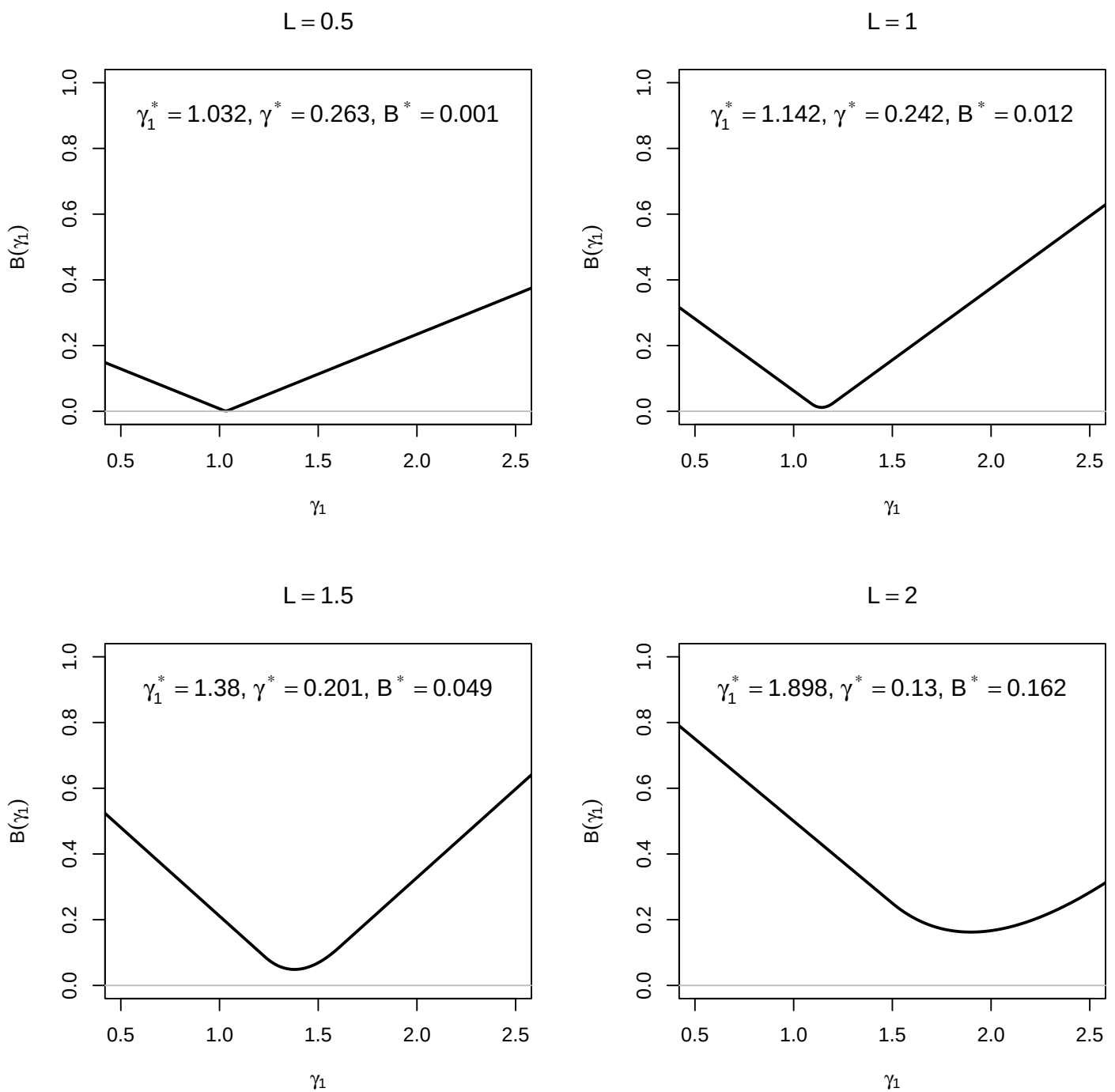


Figure 1. Average absolute bias $B(\gamma_1)$ (see definition (21) with specification (20)) for different bounds L for a uniform distributions of item difficulties on $[-L, L]$.

3.2. A Modified TPSE

A slight modification of the TPSE is now introduced that eliminates the bias component arising when the constraint $\sum_{i=1}^I b_i = 0$ is not satisfied. The core issue is that selecting the reference item such that the first item difficulty $b_1 = 0$ is essential for ensuring compliance with this sum constraint.

The modified TPSE $\tilde{b}_i$ is defined as

$$\tilde{b}_i = \hat{b}_i - \frac{1}{I} \sum_{j=1}^I \hat{b}_j, \quad (26)$$

where the estimated item difficulties $\hat{b}_i$ correspond to those obtained from the original TPSE as defined in (11). Using (26) yields

$$\tilde{b}_i = \frac{1}{I} \sum_{h=1}^I (\hat{\delta}_{ih} - \hat{\delta}_{1h}) - \frac{1}{I} \sum_{j=1}^I \frac{1}{I} \sum_{h=1}^I (\hat{\delta}_{jh} - \hat{\delta}_{1h}) = \frac{1}{I} \sum_{h=1}^I \left[\hat{\delta}_{ih} - \frac{1}{I} \sum_{j=1}^I \hat{\delta}_{jh} \right]. \quad (27)$$

This modified TPSE guarantees $\sum_{i=1}^I \tilde{b}_i = 0$ and removes the bias component introduced by violating the sum constraint under an arbitrary choice of the item-difficulty metric.

3.3. Summary

The analytical derivation shows that item difficulties in the Rasch model can be biased when applying the TPSE, even if optimal values of the nuisance parameter γ (and equivalently γ_1) are selected. Choosing an item with average item difficulty as the reference item, with $b_1 = 0$, is essential to ensure that the average item difficulty equals 0, a condition that is preferable for achieving the least biased estimates. In addition, the appropriate value of γ depends critically on the distribution of item difficulties. For more dispersed distributions, smaller values of γ are required, and the resulting TPSE-based item difficulty estimates exhibit greater bias. Importantly, no universal value of γ can yield unbiased item difficulty estimates for all possible values of the true item difficulties. This limitation reduces the appeal of the TPSE for use in empirical research.

4. Numerical Illustration

In this Numerical Illustration, the performance of the TPSE under essentially infinite sample sizes is examined. This setup allows an investigation of whether TPSE estimates are consistent.

4.1. Method

The Rasch model served as the data-generating model for examining the TPSE. The ability variable θ was assumed to follow a normal distribution with mean μ and standard deviation σ . The mean was set to 0, 0.5, or 1.0, and the standard deviation was set to either 1 or 0.0001. The latter choice resulted in a distribution concentrated at the point $\theta = \mu$.

The simulation study included $I = 10$ items. The average item difficulty $\Delta = I^{-1} \sum_{i=1}^I b_i$ was varied as 0, 0.4, and 0.8. In the condition $\Delta = 0$, the item difficulties b_i were specified as 0.000, −1.500, −1.125, −0.750, −0.375, 0.000, 0.375, 0.750, 1.125, and 1.500. For the conditions $\Delta = 0.4$ and $\Delta = 0.8$, the item difficulties for all items except the first were obtained by adding $I(I-1)^{-1}\Delta$ to the previously listed values to ensure the desired average difficulty, with the first item difficulty fixed at $b_1 = 0$.

The sample size was fixed at $N = 500,000$, representing an approximation to the population. This setup renders the simulation essentially a numerical experiment for assessing bias of the TPSE under (almost) infinite sample sizes.

In each of the 3 (mean μ) $\times 2$ (standard deviation σ) $\times 3$ (average item difficulty Δ) = 18 simulation conditions, 3000 replications were conducted. Using this large number of replications and the very large sample size results in negligible Monte Carlo error in the simulation. The TPSE was applied for γ values between 0.005 and 0.495 in steps of 0.005. For all γ values, the bias of the item parameter estimates $\hat{b}_i$ was computed. Consistent with the data-generating model, the first item difficulty $\hat{b}_1$ was fixed at 0, so the first item served as the reference item. To summarize bias across items, the average absolute bias in the estimated item difficulties was calculated. An optimal γ value, denoted by γ^* , was determined separately for each simulation condition by selecting γ such that the average absolute bias in the estimated item difficulties was minimized. Consequently, the TPSE parameter estimates were obtained under a highly optimistic best-case scenario in which sampling variability does not affect the parameter choice. Under empirical conditions, a data-driven choice of γ cannot be expected to improve the performance of the TPSE.

All analyses in this Numerical Illustration were carried out in the statistical software R (Version 4.4.1; [26]). A custom R function was used to compute the TPSE. Replication material for this Numerical Illustration is available at <https://osf.io/ef8nm> (accessed on 21 November 2025).

4.2. Results

The results of the TPSE item parameter estimates for this Numerical Illustration are now presented.

Figure 2 displays the average absolute bias in item difficulties as a function of γ , the mean μ , and the average item difficulty Δ , for a standard deviation σ of 1. The average absolute bias was close to 0 when the average item difficulty Δ equaled 0. The dependence on the mean μ was negligible. An optimal value of $\gamma^* = 0.24$ was obtained, yielding $\gamma_1 = 1.153$. The resulting average absolute bias was 0.026, which is small but still notably greater than 0.

When the average item difficulty Δ differed from 0, substantial biases occurred, and the search for an optimal γ did not yield practically unbiased estimates. These results align with the bias derivation in Section 3, which highlights the importance of fixing the reference item difficulty at 0 in such a way that the average item difficulty equals 0.

Table 1 reports the bias in item difficulties for the optimal γ parameters that minimized the average absolute bias. The estimated item difficulties show only small biases in the conditions with $\Delta = 0$. However, even in this best-case scenario, an absolute bias of 0.056 occurred for item difficulties -1.500 and 1.500 . The bias increased substantially in the conditions with $\Delta \neq 0$, where the average item difficulty differed from 0. In these situations, the TPSE has almost no practical value for empirical applications. The general pattern of results in Table 1 is essentially independent of the mean μ and the standard deviation σ of the ability variable θ .

For the best-case scenarios with $\Delta = 0$, Table 1 shows positive and negative biases for negative item difficulties, suggesting at first glance that no consistent pattern exists. This behavior can be understood by examining the bias function B (see its definition in (15)) presented in Section 3. This function is central to the TPSE because it characterizes the bias in differences of estimated item difficulties.

Figure 3 displays the function B for the optimal value $\gamma = 0.24$. The curve oscillates, with negative bias for slightly negative values of x and positive bias for more strongly negative values of x . This pattern aligns precisely with the bias structure observed in Table 1 for the $\Delta = 0$ condition.

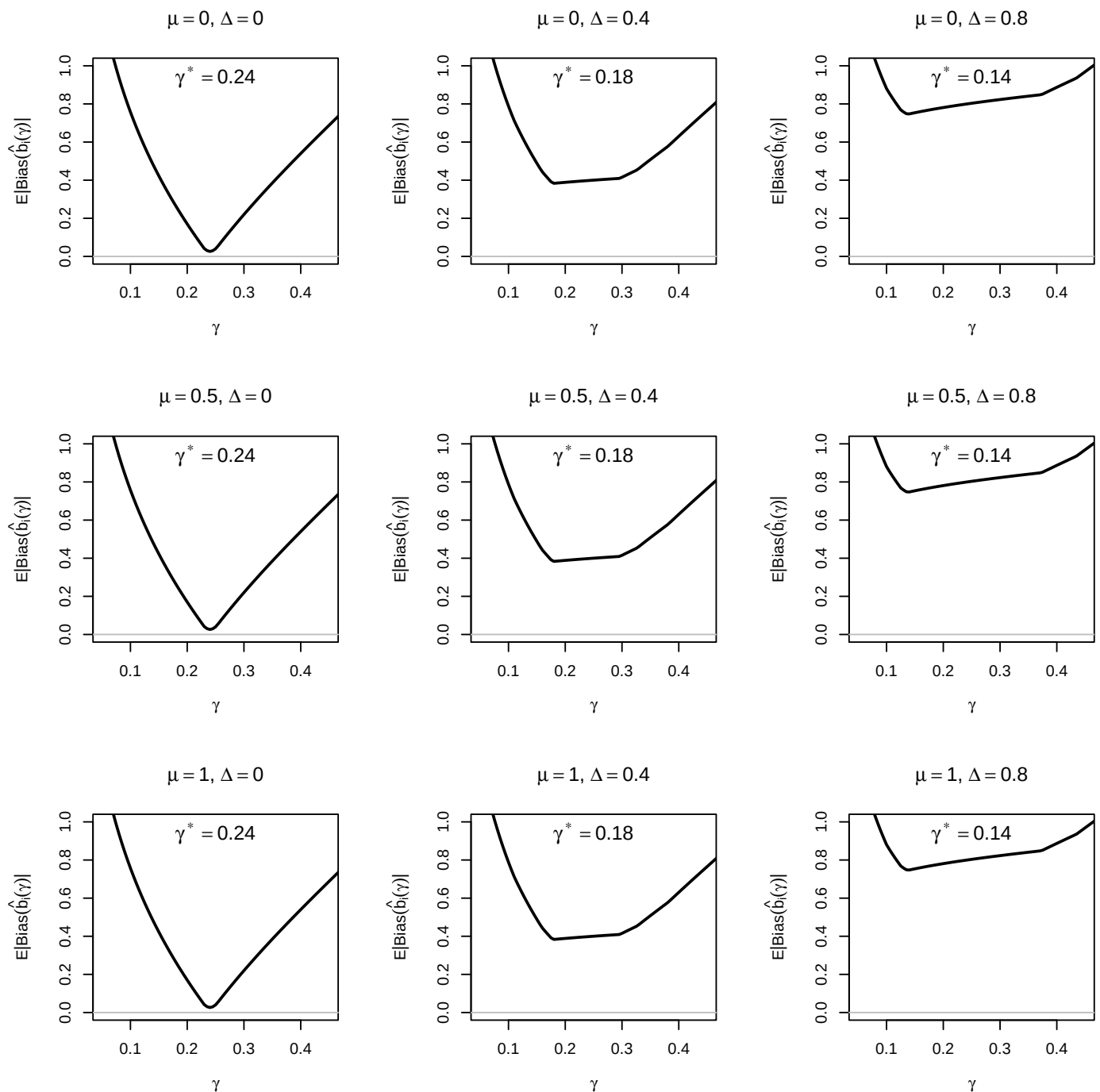


Figure 2. Numerical Illustration: Average absolute bias of estimated item difficulties as a function of γ for different means μ , true average item difficulties Δ for $\sigma = 1$. The minimal γ value is denoted by γ^* .

Table 1. Numerical Illustration: Bias in estimated item difficulties $\hat{b}_i$ using optimal γ values as a function of mean μ , true average item difficulties Δ , and standard deviation σ .

σ	β_i	$\Delta = 0, \mu =$			$\Delta = 0.4, \mu =$			$\Delta = 0.8, \mu =$		
		0	0.5	1	0	0.5	1	0	0.5	1
1	−1.500	0.056	0.056	0.056	−0.792	−0.793	−0.793	−1.526	−1.526	−1.526
	−1.125	−0.008	−0.008	−0.008	−0.748	−0.748	−0.747	−1.389	−1.389	−1.389
	−0.750	−0.030	−0.030	−0.031	−0.651	−0.651	−0.651	−1.195	−1.195	−1.195
	−0.375	−0.023	−0.023	−0.023	−0.519	−0.519	−0.519	−0.965	−0.965	−0.965
	0.000	0.000	0.000	0.000	−0.371	−0.370	−0.370	−0.721	−0.721	−0.721
	0.375	0.023	0.023	0.023	−0.227	−0.227	−0.227	−0.486	−0.486	−0.486
	0.750	0.031	0.030	0.030	−0.108	−0.108	−0.108	−0.282	−0.282	−0.282
	1.125	0.007	0.008	0.008	−0.030	−0.030	−0.030	−0.128	−0.128	−0.128
	1.500	−0.056	−0.056	−0.056	−0.006	−0.006	−0.006	−0.036	−0.036	−0.036
0.0001	−1.500	0.056	0.056	0.056	−0.793	−0.792	−0.793	−1.526	−1.526	−1.526
	−1.125	−0.008	−0.007	−0.008	−0.748	−0.748	−0.747	−1.389	−1.389	−1.389
	−0.750	−0.030	−0.030	−0.030	−0.651	−0.651	−0.651	−1.195	−1.195	−1.196
	−0.375	−0.023	−0.023	−0.023	−0.519	−0.519	−0.519	−0.965	−0.965	−0.965
	0.000	0.000	0.000	0.000	−0.370	−0.370	−0.370	−0.721	−0.721	−0.721
	0.375	0.023	0.023	0.023	−0.227	−0.227	−0.227	−0.486	−0.486	−0.486
	0.750	0.030	0.030	0.030	−0.108	−0.108	−0.108	−0.282	−0.282	−0.282
	1.125	0.008	0.008	0.008	−0.030	−0.030	−0.030	−0.128	−0.128	−0.128
	1.500	−0.056	−0.056	−0.056	−0.006	−0.006	−0.006	−0.035	−0.036	−0.035

Note. β_i = item parameter that defines true item difficulty as $b_i = \beta_i + I(I - 1)^{-1}\Delta$.

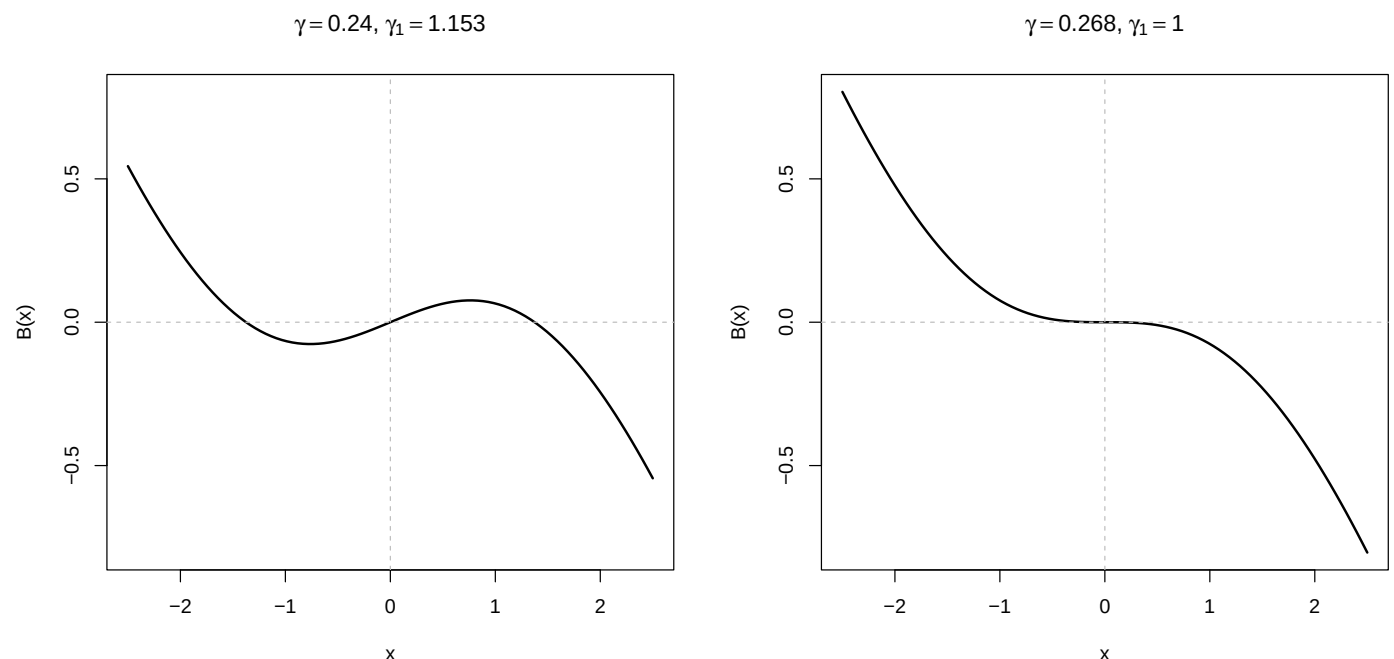


Figure 3. Bias function B (see definition in (15)) for two different γ (and γ_1) values.

5. Empirical Example

In this section, item parameter estimates from the TPSE for the Rasch model are compared with alternative parameter estimation methods using an empirical example. The dataset *MathExamp14W*, included in the R package *psychotools* (Version 0.7-5; [27]), is analyzed. It contains item response data from 729 students on 13 items from a written introductory mathematics examination (i.e., the course “Math 101”) [28].

The Rasch model was fitted using several estimation methods. Marginal maximum likelihood (MML; [9]) estimation was conducted with the `tam.mml()` function in the R

package TAM (Version 4.3-25; [29]). Conditional maximum likelihood (CML; [9]) estimation was implemented using the `immer_cml()` function in the R package `immer` (Version 1.5-13; [30,31]). Composite conditional maximum likelihood (CCML; [21,32]), another approach based on a pairwise estimation principle, was specified using the `immer_ccml()` function, also included in the R package `immer` (Version 1.5-13; [30]). The row averaging (RA; [17]) approach was implemented with the `pair()` function in the R package `pairwise` (Version 0.6.2-0; [33]). The TPSE used the item `implicit` as the reference item and was evaluated over a discrete grid of γ values ranging from 0.10 to 0.40 in increments of 0.005. All item parameters are reported using centering; that is, the item difficulties sum to zero. The optimal γ value for the TPSE was selected by minimizing the average absolute difference in item parameter estimates between the CML and TPSE methods.

Standard errors for all item parameters were obtained using a nonparametric bootstrap [34], with the standard deviation of the estimates across bootstrap samples used as the standard error estimator. The bootstrap approach was also applied to compute standard errors for differences in item parameter estimates across estimation methods [35]. These standard errors were further used to assess whether item parameters differed significantly from zero for two estimation methods. Note that the bootstrap approach automatically accounts for the fact that model parameter estimates are obtained from the same dataset, which induces dependence among the parameter estimates.

Table 2 presents the item parameter estimates together with their standard errors for the example dataset. The CML and MML estimates were very close to each other, with the exception of the eighth item (“payflow”). Notably, the correlations of item parameter estimates among the CML, MML, CCML, and RA methods all exceeded 0.9995. In contrast, the correlations between TPSE item parameter estimates and those from the other estimation methods ranged from 0.9934 to 0.9962. Although these correlations are still relatively high, they are substantially lower than the correlations observed among the other estimation methods. The optimal γ parameter for the TPSE was estimated as 0.235.

Table 2. Empirical Example: Item parameter estimates and standard errors.

Item	CML	MML	CCML	RA	TPSE	CML vs. TPSE	RA vs. TPSE	CML vs. MML
1. implicit	−0.389 (0.077)	−0.387 (0.077)	−0.347 (0.080)	−0.346 (0.080)	−0.339 (0.086)	−0.050 (0.014) *	−0.008 (0.010)	−0.002 (0.003)
2. quad	0.188 (0.079)	0.196 (0.079)	0.159 (0.075)	0.159 (0.075)	0.209 (0.079)	−0.021 (0.013)	−0.050 (0.008) *	−0.008 (0.003) *
3. deriv	−0.782 (0.094)	−0.783 (0.094)	−0.718 (0.094)	−0.709 (0.092)	−0.739 (0.099)	−0.042 (0.014) *	0.030 (0.013) *	0.001 (0.001)
4. elasticity	−1.055 (0.086)	−1.056 (0.085)	−1.048 (0.089)	−1.046 (0.088)	−1.058 (0.084)	0.003 (0.012)	0.012 (0.017)	0.001 (0.001)
5. integral	0.339 (0.078)	0.349 (0.078)	0.351 (0.078)	0.350 (0.078)	0.414 (0.080)	−0.075 (0.016) *	−0.064 (0.009) *	−0.010 (0.003) *
6. interest	−0.782 (0.086)	−0.783 (0.087)	−0.747 (0.086)	−0.745 (0.086)	−0.764 (0.088)	−0.018 (0.014)	0.019 (0.011)	0.001 (0.001)
7. annuity	−0.463 (0.079)	−0.462 (0.080)	−0.416 (0.080)	−0.417 (0.080)	−0.414 (0.087)	−0.049 (0.015) *	−0.003 (0.010)	−0.001 (0.002)
8. payflow	2.313 (0.112)	2.260 (0.102)	2.172 (0.125)	2.143 (0.125)	1.888 (0.072)	0.425 (0.056) *	0.255 (0.063) *	0.053 (0.020) *
9. matrix	−0.418 (0.075)	−0.417 (0.075)	−0.390 (0.078)	−0.394 (0.078)	−0.381 (0.083)	−0.037 (0.016) *	−0.014 (0.010)	−0.002 (0.002)
10. planning	0.763 (0.091)	0.775 (0.092)	0.739 (0.091)	0.738 (0.091)	0.806 (0.090)	−0.043 (0.019) *	−0.068 (0.013) *	−0.012 (0.003) *
11. equations	0.806 (0.085)	0.818 (0.085)	0.817 (0.087)	0.820 (0.088)	0.880 (0.082)	−0.074 (0.022) *	−0.060 (0.014) *	−0.012 (0.003) *
12. hesse	−1.271 (0.097)	−1.272 (0.096)	−1.293 (0.108)	−1.274 (0.107)	−1.291 (0.097)	0.020 (0.020)	0.016 (0.023)	0.001 (0.003)
13. lagrange	0.749 (0.089)	0.761 (0.089)	0.721 (0.084)	0.722 (0.085)	0.787 (0.083)	−0.038 (0.021)	−0.065 (0.012) *	−0.012 (0.003) *

Note. CML = conditional maximum likelihood; MML = marginal maximum likelihood; CCML = composite conditional maximum likelihood; RA = row averaging; TPSE = Tutz pairwise separation estimator; * = $p < 0.05$.

The standard errors of the TPSE item parameter estimates were, on average, slightly larger than those for CML and MML, indicating that no efficiency gains are achieved with the TPSE. The largest difference between TPSE and CML was observed for the eighth item, `payflow`, and this item parameter difference was also statistically significant. Overall, 8 out of 13 items exhibited significant parameter differences between the CML and TPSE estimates. In contrast, although the parameter differences between CML and MML were very small, the corresponding standard errors were also small, reflecting the high correlation between item parameter estimates from the two methods.

Overall, with the exception of one item, the TPSE estimates do not differ strongly from those obtained with competing estimators. Nevertheless, the question remains why the biased and inconsistent TPSE should be used in empirical research, given that it also does not reduce variance.

6. Discussion

Recently, Tutz proposed the Tutz's pairwise separation estimator (TPSE; [23]), which builds on the idea that person and item parameters can be separated in HM IRT models. However, ref. [23] provided neither a consistency proof nor a convincing large-sample simulation study demonstrating adequate statistical properties of the estimator in HM models.

This article shows that the TPSE does not offer a meaningful generalization of existing pairwise estimation methods in the Rasch model, which represents a special case of HM models. Analytical and numerical evidence for the Rasch model indicates that the TPSE will generally be biased in large samples, implying that the estimator does not yield consistent parameter estimates. The bias in estimated item difficulties depends critically on the choice of the reference item whose difficulty is fixed at 0. This reference item must be selected so that the average item difficulty equals 0; otherwise, substantial bias occurs. Even in this best-case scenario, the bias increases as the distribution of true item difficulties becomes wider. As a result, the TPSE is least suited for tests that include both very easy and very difficult items. It should be noted that, due to the inconsistency property, the usefulness of the TPSE is questionable. In the HM model, marginal maximum likelihood or Bayesian estimation of item parameters may be preferred. For the Rasch model, these estimators can also be applied or if approaches that do not require specification of the θ distribution are preferred, conditional maximum likelihood or pairwise estimation approaches [22].

It should be emphasized that this paper critiques only the use of the TPSE and not pairwise estimation approaches for the Rasch model in general. The literature has shown that consistent item parameter estimators in the Rasch model can be obtained using alternative estimators, such as those presented in the empirical example in Section 5.

This study was limited to dichotomous items, whereas [23] also considered polytomous items. Nevertheless, it is unlikely that the unsatisfactory behavior of the TPSE in the dichotomous case would fail to carry over to polytomous items.

The analytical derivation relied on a Taylor approximation up to third order for the tanh function involved in the bias of estimated item difficulties. Precision could be increased by including additional terms in the Taylor expansion. However, it remains impossible to select a universal value of the nuisance parameter γ_1 that yields unbiasedness for all possible values of the true item difficulties. The derivation presented in the present paper should therefore be viewed as a rough approximation of the bias, intended to provide insight into the general behavior of the TPSE. Nevertheless, the analytical results obtained from the approximation were supported by an Numerical Illustration that does not rely on any approximation.

Importantly, the inconsistency of the TPSE item parameter estimates also persists when the number of items tends to infinity. The TPSE continues to rely on ratios of sample proportions for item pairs, and the analytical derivations presented in this article do not depend on the number of items. As a result, the conclusions remain valid even in the limit of an infinite item set.

It cannot be expected that the bias disappears in more general HM models. The substitution of unknown item probabilities by pseudo-observations in the TPSE becomes even more critical when asymmetric link functions F are used. This conclusion follows from the behavior of the TPSE in the Rasch model for items with extreme positive or negative difficulties.

Pairwise estimation approaches are often motivated by the appealing feature that no distributional assumption for the θ variable is required, nor is a simultaneous estimation of person and item parameters. However, HM models can also be estimated using marginal maximum likelihood methods, and the θ distribution can be specified in a semiparametric manner (e.g., as in [36–39]) for a known link function F in the HM model.

Tutz [23] noted in the original paper on the TPSE that

[. . .] To this end an estimator is derived that uses pseudo observations and can be derived as a smoothing estimator. We do not claim that this is the only possible or best estimator. There might be much better ones that estimate parameters without reference to the other group of parameters.

He even concedes that the TPSE were the “best” estimator. This statement represents a notable understatement given the poor statistical performance of the TPSE in the simplest case of the Rasch model, as demonstrated in the present article. Because the TPSE performed unsatisfactorily even under the Rasch model, satisfactory properties for more complex HM models appear unlikely. Consequently, this estimator cannot be generally recommended for empirical research. Based on these findings, no plausible scenarios in empirical research can be identified in which the TPSE would be preferred over alternative estimators. It remains an open question for future research whether an estimator with satisfactory statistical properties based on the principle of parameter separation can be derived for the HM model, although I am skeptical that this is feasible.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: I would like to thank three anonymous reviewers for their valuable comments that helped improve the paper.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CCML	composite conditional maximum likelihood
CML	conditional maximum likelihood
HM	homogeneous monotone
IRF	item response function
IRT	item response theory
MML	marginal maximum likelihood
RA	row averaging
TPSE	Tutz’s pairwise separation estimator

References

1. Bock, R.D.; Gibbons, R.D. *Item Response Theory*; Wiley: Hoboken, NJ, USA, 2021. [CrossRef]
2. Chen, Y.; Li, X.; Liu, J.; Ying, Z. Item response theory—A statistical framework for educational and psychological measurement. *Stat. Sci.* **2025**, *40*, 167–194. [CrossRef]
3. Lord, F.M.; Novick, R. *Statistical Theories of Mental Test Scores*; Addison-Wesley: Reading, MA, USA, 1968.
4. Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*; Erlbaum: Hillsdale, NJ, USA, 1980. [CrossRef]
5. Tutz, G. *A Short Guide to Item Response Theory Models*; Springer: Cham, Switzerland, 2025. [CrossRef]
6. van der Linden, W.J. Unidimensional logistic response models. In *Handbook of Item Response Theory, Volume 1: Models*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 11–30.

7. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danish Institute for Educational Research: Copenhagen, Denmark, 1960.
8. Wright, B.D.; Stone, M.H. *Best Test Design*; Mesa Press: Chicago, IL, USA, 1979. Available online: <https://bit.ly/38jnLMX> (accessed on 21 November 2025).
9. Fischer, G.H.; Molenaar, I.W. (Eds.) *Rasch Models: Foundations, Recent Developments, and Applications*; Springer: New York, NY, USA, 1995. [CrossRef]
10. Linacre, J.M. Understanding Rasch measurement: Estimation methods for Rasch measures. *J. Outcome Meas.* **1999**, *3*, 382–405. Available online: <https://bit.ly/2UV6Eht> (accessed on 21 November 2025).
11. von Davier, M. The Rasch model. In *Handbook of Item Response Theory, Volume 1: Models*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 31–48.
12. Goldstein, H. Consequences of using the Rasch model for educational assessment. *Brit. Educ. Res. J.* **1979**, *5*, 211–220. [CrossRef]
13. Goldstein, H.; Wood, R. Five decades of item response modelling. *Brit. J. Math. Stat. Psychol.* **1989**, *42*, 139–167. [CrossRef]
14. Fischer, G.H. Rasch models. In *Handbook of Statistics, Vol. 26: Psychometrics*; Rao, C.R., Sinharay, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2006; pp. 515–585. [CrossRef]
15. Ballou, D. Test scaling and value-added measurement. *Educ. Financ. Policy* **2009**, *4*, 351–383. [CrossRef]
16. van der Linden, W.J. Fundamental measurement and the fundamentals of Rasch measurement. In *Objective Measurement: Theory Into Practice*; Wilson, M., Ed.; Ablex Publishing Corporation: Hillsdale, NJ, USA, 1994; Volume 2, pp. 3–24.
17. Choppin, B. A fully conditional estimation procedure for Rasch model parameters. *Eval. Educ.* **1982**, *9*, 29–42. Available online: <https://bit.ly/3gFGQ0v> (accessed on 21 November 2025).
18. Garner, M. An eigenvector method for estimating item parameters of the dichotomous and polytomous Rasch models. *J. Appl. Meas.* **2002**, *3*, 107–128. Available online: <https://bit.ly/3gKp8ZN> (accessed on 21 November 2025). [PubMed]
19. Heine, J.H.; Tarnai, C. Pairwise Rasch model item parameter recovery under sparse data conditions. *Psychol. Test Assess. Model.* **2015**, *57*, 3–36. Available online: <https://bit.ly/3sUICzC> (accessed on 21 November 2025).
20. Heine, J.H.; Robitzsch, A. Evaluating the effects of analytical decisions in large-scale assessments: Analyzing PISA mathematics 2003–2012. *Large-Scale Assess. Educ.* **2022**, *10*, 10. [CrossRef]
21. Zwinderman, A.H. Pairwise parameter estimation in Rasch models. *Appl. Psychol. Meas.* **1995**, *19*, 369–375. [CrossRef]
22. Robitzsch, A. A comprehensive simulation study of estimation methods for the Rasch model. *Stats* **2021**, *4*, 814–836. [CrossRef]
23. Tutz, G. Invariance of comparisons: Separation of item and person parameters beyond Rasch models. *J. Math. Psychol.* **2024**, *122*, 102876. [CrossRef]
24. Tutz, G. Invariance of comparisons: Separation of item and person parameters beyond Rasch models. *arXiv* **2023**, arXiv:2301.03048. [CrossRef]
25. Holland, P.W. On the sampling theory foundations of item response theory models. *Psychometrika* **1990**, *55*, 577–601. [CrossRef]
26. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2024. Available online: <https://www.R-project.org> (accessed on 15 June 2024).
27. Zeileis, A.; Strobl, C.; Wickelmaier, F.; Komboz, B.; Kopf, J.; Schneider, L.; Debelak, R. psychotools: Psychometric Modeling Infrastructure, 2025. R Package Version 0.7-5. Available online: <https://cran.r-project.org/web/packages/psychotools/index.html> (accessed on 21 October 2025). [CrossRef]
28. Zeileis, A. Examining exams using Rasch models and assessment of measurement invariance. *Austrian J. Stat.* **2025**, *54*, 9–26. [CrossRef]
29. Robitzsch, A.; Kiefer, T.; Wu, M. TAM: Test Analysis Modules, 2025. R Package Version 4.3-25. Available online: <https://cran.r-project.org/web/packages/TAM/index.html> (accessed on 28 August 2025). [CrossRef]
30. Robitzsch, A.; Steinfeld, J. immer: Item Response Models for Multiple Ratings, 2024. R Package Version 1.5-13. Available online: <https://cran.r-project.org/web/packages/immer/index.html> (accessed on 21 March 2024). [CrossRef]
31. Robitzsch, A.; Steinfeld, J. Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychol. Test Assess. Model.* **2018**, *60*, 101–138. Available online: <https://bit.ly/3mFnn3U> (accessed on 21 November 2025).
32. van der Linden, W.J.; Eggen, T.J.H.M. An empirical Bayesian approach to item banking. *Appl. Psychol. Meas.* **1986**, *10*, 345–354. [CrossRef]
33. Heine, J.H. pairwise: Rasch Model Parameters by Pairwise Algorithm, 2025. R Package Version 0.6.2-0. Available online: <https://cran.r-project.org/web/packages/pairwise/index.html> (accessed on 12 September 2025). [CrossRef]
34. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; CRC Press: Boca Raton, FL, USA, 1994. [CrossRef]
35. Robitzsch, A.; Lüdtke, O. Mean comparisons of many groups in the presence of DIF: An evaluation of linking and concurrent scaling approaches. *J. Educ. Behav. Stat.* **2022**, *47*, 36–68. [CrossRef]
36. De Leeuw, J.; Verhelst, N. Maximum likelihood estimation in generalized Rasch models. *J. Educ. Behav. Stat.* **1986**, *11*, 183–196. [CrossRef]

37. Bartolucci, F. A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika* **2007**, *72*, 141–157. [[CrossRef](#)]
38. Lindsay, B.; Clogg, C.C.; Grego, J. Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Am. Stat. Assoc.* **1991**, *86*, 96–107. [[CrossRef](#)]
39. von Davier, M. A general diagnostic model applied to language testing data. *Brit. J. Math. Stat. Psychol.* **2008**, *61*, 287–307. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.