psych

MDPI

Article

Item Parameter Estimation in Multistage Designs: A Comparison of Different Estimation Approaches for the Rasch Model

Jan Steinfeld [1,2,*] and Alexander Robitzsch [3,4]

[1] Differential Psychology and Psychological Assessment, Department of Developmental and Educational Psychology, Faculty of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria
[2] Austrian Federal Ministry of Education, Science and Research, A-1010 Vienna, Austria
[3] IPN—Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, D-24118 Kiel, Germany; robitzsch@leibniz-ipn.de
[4] Center for International Student Assessment (ZIB), D-80333 München, Germany
* Correspondence: jan.steinfeld@univie.ac.at

Abstract: There is some debate in the psychometric literature about item parameter estimation in multistage designs. It is occasionally argued that the conditional maximum likelihood (CML) method is superior to the marginal maximum likelihood method (MML) because no assumptions have to be made about the trait distribution. However, CML estimation in its original formulation leads to biased item parameter estimates. Zwitser and Maris (2015, *Psychometrika*) proposed a modified conditional maximum likelihood estimation method for multistage designs that provides practically unbiased item parameter estimates. In this article, the differences between different estimation approaches for multistage designs were investigated in a simulation study. Four different estimation conditions (CML, CML estimation with the consideration of the respective MST design, MML with the assumption of a normal distribution, and MML with log-linear smoothing) were examined using a simulation study, considering different multistage designs, number of items, sample size, and trait distributions. The results showed that in the case of the substantial violation of the normal distribution, the CML method seemed to be preferable to MML estimation employing a misspecified normal trait distribution, especially if the number of items and sample size increased. However, MML estimation using log-linear smoothing lea to results that were very similar to the CML method with the consideration of the respective MST design.

Keywords: multistage testing; Rasch model; marginal maximum likelihood; conditional maximum likelihood; parameter estimation; log-linear smoothing

check for updates

Citation: Steinfeld, J.; Robitzsch, A. Item Parameter Estimation in Multistage Designs: A Comparison of Different Estimation Approaches for the Rasch Model. *Psych* **2021**, *3*, 279–307. https://doi.org/10.3390/psych3030022

Academic Editor: Holmes Finch

Received: 30 April 2021
Accepted: 2 July 2021
Published: 8 July 2021

1. Introduction

Several studies have shown adaptive test designs such as computerized adaptive tests (CATs; [1–7]) or multistage tests (MST; [8–13]) are usually more efficient in terms of shorter test lengths, providing equal or even higher measurement precision and higher predictive validity, compared to linear fixed-length tests (LFTs; [6,7,14–23]). The advantages of adaptive tests are particularly evident for more extreme abilities at the lower and upper end of the measurement scale [6,15,24].

In situations with administration time constraints, CATs can be a good choice and should be considered. However, a decision in favor of adaptive tests also means that some disadvantages are taken for granted. Some will be explained in the following. It should become clear that MST designs, compared to CATs, do not share many of these disadvantages, which has probably also led to its popularity and use in educational measurement and, in particular, international large-scale assessments (ILSAs; e.g., [16]). In recent years, several well-known programs, such as the Programme for International Student Assessment (PISA; [25]), the Programme for The International Assessment of Adult

Competencies (PIAAC; [26]), Trends in the International Mathematics and Science Study collection cycle 2019 on computer-based assessment systems (eTIMSS, TIMSS; [27]), or the National Assessment of Educational Progress (NAEP; [28,29]), applied MST designs and might have contributed to its popularity. Besides ILSAs, there are several other areas with successful applications in the past decade, such as psychological assessment (e.g., [30]), or classroom assessments [16]. It can be summarized that the application of adaptive testing currently has become an essential testing method (e.g., [31,32]).

In the following, we refer to MSTs and CATs in their more classical form, even if some contributions do not separate both designs so strictly from one another. Chang [16], for example, stated that both designs could be regarded as sequential designs (see also [33–36] for dynamic multistage designs).

Here, CATs should therefore be understood as adaptive designs on the item level. Based on one or more item selection algorithms, the best-suited item is selected. The maximum of information is often defined with a success rate of 50% for this item. If the item pool is large enough for the desired measurement accuracy, the smallest number of items is required in CATs. Therefore, the efficiency is theoretically the largest if the item pool is large enough. Some indices to measure the amount of adaptation in practice were recently discussed by Wyse and McBride [37].

In MSTs, the decision points are modules. These are collections of items with mostly related content (see also the comparison to testlets; [6,38]), certain mean item difficulties, and variances. At the start, test persons receive a routing module and, based on the performance in this module and performance-related prior information, if available, one or more additional modules. Each additional module in this routing process describes a stage in the MST design. Each stage consists of at least two modules (see Figure 1 for an example). The specific combination of processed modules in the routing process is called a path. Different groups of modules, stages, and paths are called panels. Panels can be seen as parallel forms in LFTs. Routing in the MST context is branching from one to the next module, based on pre-specified rules.

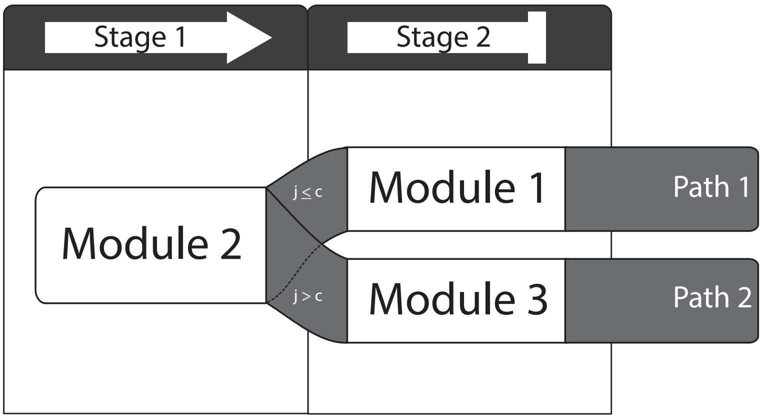

Figure 1. Example of a multistage design with three modules, two stages, and two paths. Note: j = score in Module 2; c = given cutoff value.

As with all adaptive designs, the selection of items or modules is the central part of the design, and much research has been performed to serve different needs. In particular, in CATs, the item selection can become very complex. Additional considerations can refer to, e.g., content balancing or strategies to avoid overexposure and/or underexposure of items. Next to the desired purpose of that algorithm, there might also be some disadvantages, which can negatively impact the validity and the fairness of the test. In particular, in CATs, the item exposure control can become a challenging task [13,39].

Overexposure might be a problem if the information of those items processed more often is shared across test persons. This can threaten the validity of the test because the performance of the test persons can no longer be separable between ability and knowledge.

Especially with high-stakes tests, it might be a major problem, where industry could quickly build up to collect the information of items [40]. While simply increasing the item pool is not the solution [39], additional algorithms must be considered. Concerning underexposure, economic considerations are probably more in the foreground, as the construction of items is very expensive. However, this can also lead to problems in parameter estimation if the sample size per item is low, which subsequently results in the inaccurate estimation of item parameters and the standard errors. Here, MST designs seem to show their advantages, as they can be designed and checked before they are applied. Hence, no additional algorithms are necessary.

1.1. Motivation

An essential factor in every test is the motivation of the test persons (see, e.g., [41–43]). It has been reported that due to the better match of the item difficulty and the person's ability, test persons, especially those with low abilities, are more motivated to proceed, sometimes less bored, and more committed during the test [44–50]. On the other hand, there are several contributions concerning CATs that report negative psychological effects of the demanding item selection. Kimura [51] stated that this could lead to negative test experience,s as well as lower motivation, lower self-confidence, and increased test anxiety (see also [52–58]). These psychological variables seem to be an important topic in testing since they could negatively affect the persons' test performance [56,59,60]. Motivation is a key factor in every low-stakes test such as ILSAs since unmotivated participants might influence the test results and thus the validity of the test (see, e.g., [61]). It seems to be central and can be deduced from these contributions that the impact on motivation or boredom, but also anxiety, should not be ignored, as this can significantly influence the test results [62,63]. Finally, this contributes to standardization and thus to reliable results and more valid parameter estimates [64,65].

MST designs are conceptualized before the actual application. The items are explicitly assigned to modules, and every path of that design can be reviewed in advance. Therefore, these mentioned aspects can be verified before the application, and no additional algorithms are required during the actual application.

1.2. Test Anxiety

Increased test anxiety among test participants is another reported psychological effect in CATs [60]. Due to the lack of the possibility to review items that have already been processed and, if necessary, changed by the test person, test anxiety might also be further increased [66–70]. An item revision in CATs is not possible [7,71,72], because the item selection in CATs is based on the responses already given. Hence, changing responses retrospectively may impact the measurement precision, which results in larger standard errors [69,73–78]. Therefore, allowing item revision within CATs has been controversially discussed in the literature, even if some contributions encountered this measurement problem (see, e.g., [66,77,79–82]). While it can be argued that only a few persons might change their responses [83], a lack of this ability appears to contribute to increased test anxiety. However, it is also reported that subsequent changes to given responses are mostly from wrong to processed correct [83] and thus not only affect the psychological aspects, but also the validity of the test scores.

Several studies suggested methods to allow a (limited) item review in CATs while avoiding the negative effects of the lower measurement accuracy or the extension of the test at the same time [68,75,77,81,82,84]. However, the proposals can also be viewed critically. For example, Zwick and Bridgeman [85] found that more experienced test persons may use the review options more often than others. This could again harm the validity of the test, while the absence of the item review affects all persons across the entire skill range equally [60]. Next to the possibility of reprocessing the responses in CATs, this option can also be used to manipulate the test score [84,86]. Wainer [76] described one of these strategies, in which a test person first gives only incorrect responses to continuously obtain easier items. At the end of the test, all given responses are then corrected, which results in large measurement errors. Kingsbury [87] described a strategy in which test persons recognize whether a subsequent item is easier or more difficult than the one they have just worked on and obtain information about the given response. If the following item is easier, which hints that the prior response might be wrong, the response can be changed on this item; see also [88]. In MSTs, all test persons have the same chance to review their given responses and change them before taking on a new module. It is, therefore, to be expected that test anxiety will be lower with MSTs.

1.3. Routing in Adaptive Designs

Item selection algorithms are one of the key factors in CATs, especially when it comes to maximizing the test economy and thus shortening the test length [16]. Increasing the test efficiency can also be viewed critically, as we will discuss later. When choosing one of the selection algorithms, the optimization and the associated negative effects should be considered. Furthermore, the item selection is also related to considerations regarding under- and over-exposure, as well as considerations of the safety aspects. Some selection algorithms can be found in Chang [16].

In this context, deterministic means that persons with the same performance in the same module $\mathbf{m}^{[\mathbf{b}]}$ of B modules with $b = (1, \ldots, B)$ in the same stage are routed to the same subsequent module. A decision base can be, e.g., the number of solved items (number-correct score; NC). Assuming a person θ_p achieves a score j in the module $\mathbf{m}^{[\mathbf{b}]}$, this person, given a cutoff value c, is routed to an easier module in the cases $j < c$ or $j \leq c$ (that is, once again, performed deterministically by the test author) and, in the remaining cases, a more difficult module (see also [6,12]). In this simple case, the decision to route from one module to the next is only made based on the performance in the module currently being processed. This can easily be expanded by including the information from all previously processed modules in the decision. This type of routing should be referred to as the cumulative number-correct score (cNC; [89,90]). Since the information about the persons' ability across modules is used, theoretically, a more valid routing is possible. In addition to the raw scores, the routing decision can also be made based on specifically processed items. Since item parameters are known, person parameters can be estimated a priori via the respective item combinations. This type of routing is referred to in the literature as item response theory (IRT)-based routing [91]. The decision for a routing strategy in MST is linked to the efficiency of the proposed design and can also impact the precision of item parameter estimation [6]. The available strategies can roughly be grouped into deterministic and probabilistic ones. Svetina et al. [89] compared different routing strategies. The authors concluded that the IRT-based routing performed best, but the NC-based routing was not significantly worse when it came to the median of person parameter recovery rates. An additional argument for NC-based routing is that it is much easier to implement.

In the mentioned probabilistic routing, the routing rule $j < c$, respectively $j \leq c$, is expanded with an additional probability based on the performance j. This means that routing into an easier module is not solely based on the cutoff value c, but rather with a previously defined probability p, depending on the individual score j of person p. With the counter-probability $1 - \mathsf{p}$ and the same score j, the person is routed to a more difficult module. This type of routing is used, for example, in the PIAAC [32,92,93]. In addition to the deterministic definition of the cutoff values c, additional thresholds are defined for each decision stage and score.

A motivation to use probabilistic routing instead of exclusively deterministic is the possibility of being able to better control the exposure rate so that it is ensured across all proficiency levels that a minimum number of sufficient responses per item is guaranteed, even with difficult tasks (see, e.g., [32,93]).

To summarize: MSTs can be seen as a design with advantages from two perspectives. There are fully adaptive item-by-item designs such as CATs with a very high test economy [14,23,94], on the one hand, and LFTs, on the other [94]. MSTs allow for more efficient testing; test persons can review items within modules they have already worked on and change their responses if necessary. The design can be examined by the test authors concerning the item content regarding content balancing and security concerns, but also possible differential item functioning. Even overexposure and underexposure can be controlled more easily [95]. While CATs are tied to the computer, MSTs can also be administered as paper-pencil tests [19,22,30].

2. Item Parameter Estimation

Item parameter estimation in adaptive designs is an important topic and relates to the MST's main component of this contribution. For the calibration of an item pool, with data obtained by an MST, an item response theory model such as the Rasch model (1PL; [96]) is fitted. Item parameters are typically regarded as fixed, and persons are treated as either fixed or random (see, e.g., [9,97–100], for a further discussion on this topic). Several methods are available, which will be briefly discussed in the following.

These are the marginal maximum likelihood method (MML; [101–103]) and the conditional maximum likelihood method (CML; [104,105]). Various considerations can lead to choosing one of these estimation methods, such as the flexibility of that approach or more fundamental beliefs about the method.

The MML estimation method can also be applied in MST designs without leading to biased item parameter estimates (see, e.g., [106–108]). The CML-based parameter estimation in MSTs, without severely biased item parameter estimates [108], is only feasible by modifying the CML estimation method proposed by Zwitser and Maris [109]. Besides the relatively newly proposed modification of the CML approach, the normal MML method and models with non-normal trait [110] are available. It is frequently argued that the CML estimation method enables the estimation of item parameters independent of the distribution assumptions of the trait [107–109,111]. Comparisons between CML and MML estimation in MSTs showed biased item parameter estimates in MML if the distribution assumption deviates severely from the true distribution (see, e.g., [109]). In our contribution, the estimation methods were systematically examined and compared. In this context, it seems very interesting that scaling the data using a multigroup model, in which the groups are represented by the respective paths in the MST design, seems to lead to severely biased parameter estimates [106].

In the following, we only considered dichotomous item responses and utilized the 1PL model. In the 1PL model, the probability of solving item i with difficulty β_i by person p with ability θ_p can be expressed as:

$$\mathrm{P}(X_{pi} = x_{pi} \mid \theta_p, \beta_i) = \frac{\exp[x_{pi}(\theta_p - \beta_i)]}{1 + \exp(\theta_p - \beta_i)}, \tag{1}$$

with $x_{pi} = 1$. Then, the likelihood $L(\mathbf{x}_p \mid \theta_p, \boldsymbol{\beta})$ with responses $\mathbf{x}_p = (x_{p1}, x_{p2}, \ldots, x_{pI})$ of the test person p with ability θ_p and the item difficulty $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_I)$ can be expressed as follows:

$$L(\mathbf{x}_p \mid \theta_p, \boldsymbol{\beta}) = \frac{\exp(r_p\theta_p - \sum_{i=1}^{I} x_{pi}\beta_i)}{\prod_{i=1}^{I}(1 + \exp(\theta_p - \beta_i))} \tag{2}$$

with r_p as the raw score of person p with $r_p = \sum_{i=1}^{I} x_{pi}$. Equation (2) can be seen as the starting point for the following approaches in parameter estimation. The likelihood for the response matrix $\mathbf{X}$ can be expressed as:

$$L(\mathbf{X} \mid \theta, \boldsymbol{\beta}) = \frac{\exp(\sum_{p=1}^{P} r_p\theta_p - \sum_{p=1}^{P}\sum_{i=1}^{I} x_{pi}\beta_i)}{\prod_{p=1}^{P}\prod_{i=1}^{I}(1 + \exp(\theta_p - \beta_i))} \tag{3}$$

2.1. Marginal Maximum Likelihood Estimation

For the estimation in the parametric case (see Equation (4)), a distribution G with probability density function $g(\theta;\boldsymbol{\alpha})$ with a vector $\boldsymbol{\alpha}$ containing the parameters of the latent ability distribution is introduced for person parameter θ. It is assumed that the persons are a random sample from this population, e.g., $\theta \sim \mathrm{N}(\mu,\, \sigma^2)$. The random variable θ is integrated out of the marginal log-likelihood function. For parameter estimation in MST designs, Glas [108] and Zwitser and Maris [109] stated that the distributional assumptions could be incorrect, and the estimated item parameter estimates can be severely biased. Therefore, the following simulation should shed some light on this.

Data collected based on the MST design have missing values due to the design. Mislevy and Sheehan [112], referring to Rubin [113], showed that MML provides consistent estimates in incomplete designs in general (see also [106]). For MST designs, it can be shown that MML can also be applied to MST, following this justification [106,109]. Based on the likelihood function (3), in the MML case, the likelihood for the observed data matrix $\mathbf{X}$ is the product of the integrals of the respective likelihood of the response patterns $\mathbf{x}_i$.

$$L_{MML}(\mathbf{X} \mid \boldsymbol{\beta}, \mu, \sigma^2) = \prod_{r=0}^{I} \left[\int \frac{\exp(r\theta - \sum_{i=1}^{I} s_i\beta_i)}{\prod_{i=1}^{I}(1+\exp(\theta-\beta_i))} g(\theta;\boldsymbol{\alpha}) d\theta \right]^{n_r} \tag{4}$$

with $s_i = \sum_{p=1}^{P} x_{pi}$ the item score of item i, n_r as the number of test persons with the raw score r, and $\boldsymbol{\alpha}$ as a parameters for the distribution G.

For model identification purposes, if a normal distribution is assumed, the mean is fixed to zero $\mu = 0$, and σ^2 is freely estimated. Therefore, the marginal likelihood is no longer dependent on θ (see Equation (4)). The integral in Equation (4) can be solved by, e.g., Gauss–Hermite quadrature by summing over a finite number of discrete quadrature points θ_q with $q = (1, \cdots, Q)$ and the corresponding weights $w_q = w_q$ (see, e.g., [101,102]).

$$L_{MML}(\mathbf{X} \mid \boldsymbol{\beta}, G) = \prod_{r=0}^{I} \left[\exp(-\sum_{i=1}^{I} s_i\beta_i) \sum_{q=1}^{Q} \left(\frac{\exp(r\theta_q)}{\prod_{i=1}^{I}(1+\exp(\theta_q-\beta_i))} \right) w_q \right]^{n_r} \tag{5}$$

Marginal Maximum Likelihood with Log-Linear Smoothing

For the specification of the unknown latent ability distribution G in Equation (4), both parametric and nonparametric strategies are available. Another interesting approach for the specification, which is flexible and parsimonious in terms of the number of parameters to be estimated, is the application of log-linear smoothing (LLS; [110,114,115]). In IRT, this method was used, for example, by Xu and von Davier [110]. They fitted an unsaturated log-linear model in the framework of a general diagnostic model (GDM; [116]) to determine the discrete (latent) ability distribution $g(\theta)$. The LLS model used here in the case of the 1PL can be described as $\log w_q \cong \delta_0 + \sum_{m=1}^{M} \delta_m w_q^m$ [115,117]. Here, $\log w_q$ describes the logarithmic weighted quadrature points $(\theta_1, \cdots, \theta_Q)$. The intercept δ_0 is a normalization constant, M the moments to be fitted, and δ_m the dependent coefficients to be estimated. The central property of log-linear smoothing is the matching of the moments of the empirical distribution.

An interesting connection between the MML parameters' estimation outlined above in Section 2.1 using a nonparametric approach as described by Bock and Aitkin [101] (also referred to as a Bock–Aitkin or the empirical histogram (EH) solution) and the LLS is that the former can be seen as a special case of the LLS method with $M = Q - 1$ moments.

The LLS is integrated into the EM algorithm [110] to estimate $\boldsymbol{\beta}$ since the number of expected persons (expected frequencies) at each quadrature point g_q is unobserved. An LLS with $M = 2$ moments is equivalent to a discretized (standard) normal distribution (exactly two parameters are necessary, μ and σ^2) (see [117]). The specification of more than two moments allows, e.g., the specification of skewed latent variables [118].

Casabianca and Lewis [115] showed in detailed and promising simulation studies that the LLS method leads to better parameter recovery if the specified distribution deviates from the true empirical ones. By specifying up to four moments, bimodal distributions could be captured. It is also worth mentioning that there may be less effort for users to use this method since only the number of moments has to be specified.

2.2. Conditional Maximum Likelihood Estimation

Unlike the MML method, CML does not require assumptions for the distribution of the traits. Here, the person parameter is eliminated from the likelihood due to conditioning on the raw scores r_p, which is referred to as *minimal sufficient statistic* for person parameter θ_p [96,104,105,119] in Equation (6). Therefore, only item parameters β_i, but no person parameter θ_p, are estimated, which have to be determined afterwards. In the following, the likelihood for the response matrix $\mathbf{X}$ in the CML case is outlined following Equation (3) again.

For the estimation of item parameter, the calculation of the elementary symmetric function (ESF) $\gamma(r, \boldsymbol{\beta})$ of order r_p of $\beta_1, \beta_2, \ldots, \beta_I$ is the crucial part of the likelihood in CML. Different methods have been proposed, which differ mainly in accuracy and speed [120–122].

There are $\binom{I}{r_p}$ different possibilities to obtain the score r_p for a person with the ability θ_p. The sum over these different possibilities results in $\gamma(r, \boldsymbol{\beta}) = \sum_{\mathrm{x}_i | \sum \mathrm{x}_i = r} \exp(-\sum_{i=1}^{I} x_i \beta_i)$, with given item difficulty β_i, as well as the responses x_i for a given score r.

$$L_{CML}(\mathbf{X} \mid \mathbf{r}, \boldsymbol{\beta}) = \frac{L(\mathbf{X} \mid \boldsymbol{\theta}, \boldsymbol{\beta})}{L(\mathbf{r} \mid \boldsymbol{\beta})} \tag{6}$$

The likelihood of the response vector $\mathbf{r}$ can be written as:

$$L(\mathbf{r} \mid \boldsymbol{\beta}) = \frac{\exp(\sum_{p=1}^{P} r_p \theta_p) \prod_{p=1}^{P} \sum_{\mathrm{x}_{pi}}^{r_p} \exp(-\sum_{i=1}^{I} x_{pi} \beta_i)}{\prod_{p=1}^{P} \prod_{i=1}^{I} (1 + \exp(\theta_p - \beta_i))} \tag{7}$$

The likelihood in Equation (6) can then be written using Equations (3) and (7) in the CML case as follows:

$$L_{CML}(\mathbf{X} \mid \mathbf{r}, \boldsymbol{\beta}) = \frac{\exp(-\sum_{i=1}^{I} s_i \beta_i)}{\prod_{r=0}^{I} \gamma(r, \boldsymbol{\beta})^{n_r}} \tag{8}$$

The resulting estimates $\hat{\boldsymbol{\beta}}$ are consistent, asymptotically efficient, and asymptotically normally distributed [99].

CML Approach of Zwitser and Maris (2015)

Glas [108] stated that ignoring the MST design in the CML item parameter estimation process leads to severely biased estimates (see also [107,111]). Based on these results, it has long been recommended not to use the CML method for MST designs. The MML method offered an alternative, or the parameter of the items for each path or module could be estimated separately using the CML method [123]. The latter has the major disadvantage that item parameters estimated in this way can no longer be compared. Recently, this CML estimation problem could be solved for deterministic routing while considering the respective MST design in the CML estimation process [109]. To solve this problem, the symmetric function has to be modified, such that only those raw scores are considered, which can occur due to the specific MST design. This leads to consistent item parameter estimates. There are currently two R [124] packages for this method: `dexterMST` [125] and `tmt` [126]. The modified CML estimate is outlined in the following. In the deterministic case, a person with score j is routed from one module $\mathbf{m}^{[\mathbf{b}]}$ to the next module based on a cut-score c. Based on the design in Figure 1, the probability of reaching a score of $x^{[1,2]}$ in the modules $\mathbf{m}^{[1,2]}$ with ability θ, and the number of solved items in the module $\mathbf{m}^{[1]}$ being less than or equal to the cut-score c with $X_+^{[1]} \leq c$, can be described as follows:

$$P_{\mathbf{m}^{[1,2]}}(\mathbf{x}^{[1,2]} \mid \theta, X_+^{[1]} \leq c) = \frac{P_{\mathbf{m}^{[1,2]}}(\mathbf{x}^{[1,2]}, X_+^{[1]} \leq c \mid \theta)}{P_{\mathbf{m}^{[1,2]}}(X_+^{[1]} \leq c \mid \theta)} = \frac{P_{\mathbf{m}^{[1,2]}}(\mathbf{x}^{[1,2]} \mid \theta)}{P_{\mathbf{m}^{[1,2]}}(X_+^{[1]} \leq c \mid \theta)} \tag{9}$$

The ESF as described above can be written as $\gamma_s(\mathbf{m}^{[\mathbf{b}]}) = \sum_{\mathbf{x}:x_+^{[b]}=s} \prod_i \exp(-x_i^{[b]}\beta_i^{[b]})$ and rearranged as $\gamma_{x_+}(\mathbf{m}) = \sum_{i+j+k=x_+} \gamma_i(\mathbf{m}^{[1]})\gamma_j(\mathbf{m}^{[2]})\gamma_k(\mathbf{m}^{[3]})$. Here, the ESF is first evaluated for each module separately and then for a specific path of the MST design. Zwitser and Maris [109] proposed to partition the denominator of the likelihood into the sum of items $j = 0, \cdots, c$ in the first module and $x_+^{[1,2]} - j$ items in the second module. Equation (9) can be factored as:

$$P_{\mathbf{m}^{[1,2]}}(\mathbf{x}^{[1,2]} \mid \theta, X_+^{[1]} \leq c) = P_{\mathbf{m}^{[1,2]}}(\mathbf{x}^{[1,2]} \mid x_+^{[1,2]}, X_+^{[1]} \leq c) P_{\mathbf{m}^{[1,2]}}(x_+^{[1,2]} \mid \theta, X_+^{[1]} \leq c) \tag{10}$$

Inserting Equation (10) into the common CML approach results in:

$$P_{\mathbf{m}^{[1,2]}}(\mathbf{x}^{[1,2]} \mid x_+^{[1,2]}) = \frac{\prod_i \exp(-x_i^{[1]}\beta_i^{[1]}) \prod_j \exp(-x_j^{[2]}\beta_j^{[2]})}{\sum_{j=0}^{n^{[1,2]}} \gamma_j(\mathbf{m}^{[1]})\gamma_{x_+^{[1,2]}-j}(\mathbf{m}^{[2]})} \tag{11}$$

The probability of $X_+^{[1]}$ being less than or equal to c conditional on $x_+^{[1,2]}$:

$$P_{\mathbf{m}^{[1,2]}}(X_+^{[1]} \leq c \mid x_+^{[1,2]}) = \frac{\sum_{j=0}^{c} \gamma_j(\mathbf{m}^{[1]})\gamma_{x_+^{[1,2]}-j}(\mathbf{m}^{[2]})}{\sum_{j=0}^{n^{[1,2]}} \gamma_j(\mathbf{m}^{[1]})\gamma_{x_+^{[1,2]}-j}(\mathbf{m}^{[2]})} \tag{12}$$

Following Equations (11) and (12), we obtain:

$$\begin{aligned} P_{\mathbf{m}^{[1,2]}}(\mathbf{x}^{[1,2]} \mid x_+^{[1,2]}, X_+^{[1]} \leq c) &= \frac{P_{\mathbf{m}^{[1,2]}}(\mathbf{x}^{[1,2]}, X_+^{[1]} \leq c \mid x_+^{[1,2]})}{P_{\mathbf{m}^{[1,2]}}(X_+^{[1]} \leq c \mid x_+^{[1,2]})} \\ &= \frac{P_{\mathbf{m}^{[1,2]}}(\mathbf{x}^{[1,2]} \mid x_+^{[1,2]})}{P_{\mathbf{m}^{[1,2]}}(X_+^{[1]} \leq c \mid x_+^{[1,2]})} \\ &= \frac{\prod_i \exp(-x_i^{[1]}\beta_i^{[1]}) \prod_j \exp(-x_j^{[2]}\beta_j^{[2]})}{\sum_{j=0}^{c} \gamma_j(\mathbf{m}^{[1]})\gamma_{x_+^{[1,2]}-j}(\mathbf{m}^{[2]})} \end{aligned} \tag{13}$$

and further:

$$P_{\mathbf{m}^{[1,2]}}(x_+^{[1]}, x_+^{[2]} \mid \theta) = \frac{\gamma_{x_+^{[1]}}(\mathbf{m}^{[1]})\gamma_{x_+^{[2]}}(\mathbf{m}^{[2]}) \exp\left\{(x_+^{[1]} + x_+^{[2]})\theta\right\}}{\sum_{0 \leq j+k \leq n^{[1,2]}} \gamma_j(\mathbf{m}^{[1]})\gamma_k(\mathbf{m}^{[2]}) \exp\{(j+k)\theta\}} \tag{14}$$

Taking the same considerations from Equation (14) for the following:

$$
\begin{aligned}
P_{\mathbf{m}^{[1,2]}}(x_+^{[1,2]} \mid \theta, X_+^{[1]} \leq c) &= \frac{P_{\mathbf{m}^{[1,2]}}(x_+^{[1,2]}, X_+^{[1]} \leq c \mid \theta)}{P_{\mathbf{m}^{[1,2]}}(X_+^{[1]} \leq c \mid \theta)} \\
&= \frac{\sum\limits_{j \leq c} \gamma_j(\mathbf{m}^{[1]})\gamma_{x_+^{[1,2]}-j}(\mathbf{m}^{[2]}) \exp\left\{(x_+^{[1,2]})\theta\right\}}{\sum\limits_{\substack{0 \leq j+k \leq n^{[1,2]} \\ j \leq c}} \gamma_j(\mathbf{m}^{[1]})\gamma_k(\mathbf{m}^{[2]}) \exp\{(j+k)\theta\}}
\end{aligned}
\tag{15}
$$

then it follows that:

$$
P_{\mathbf{m}^{[1,2]}}(\mathbf{x}^{[1,2]} \mid \theta) = P_{\mathbf{m}^{[1]}}(\mathbf{x}^{[1]} \mid x_+^{[1]})P_{\mathbf{m}^{[2]}}(\mathbf{x}^{[2]} \mid x_+^{[2]})P_{\mathbf{m}^{[1,2]}}(x_+^{[1]}, x_+^{[2]} \mid x_+^{[1,2]})P_{\mathbf{m}^{[1,2]}}(x_+^{[1,2]} \mid \theta) \tag{16}
$$

Using Equations (13), (15) and (16), Equation (10) follows. Therefore, it can be concluded that after the integration of additional design information in the MST design, the CML item parameter estimation is justified.

3. Simulation Study

A Monte Carlo simulation was carried out to provide information on the influence of different trait distributions on the estimation of item parameters in MST designs. In addition to the different trait distributions (normal, bimodal, skewed, and χ^2 with $df = 1$), the test length (I = 15, 35, and 60 items), different MST designs, and sample sizes (N = 100, 300, 500, and 1000) were considered. All conditions were simulated as MSTs, as well as fixed-length tests. The simulation and all conditions are explained in detail below. MST designs can be expanded to more modules, items within modules, and more stages. It is important to note that, branching on the item level as is the case with CATs, CML estimation is not possible. As stated by Zwitser and Maris [109] for CATs, the information about the item parameters is bound in the design and thus not available for CML parameter estimation. Therefore, CAT designs were not considered here.

3.1. Data Generation

For all MST conditions, a two-stage design was used (see Figure 1). All MST conditions started with the routing module $\mathbf{m}^{[2]}$ and were subsequently routed in one additional module. The module with easier items was the module $\mathbf{m}^{[1]}$ and the module with more difficult items $\mathbf{m}^{[3]}$. The entire routing was based on the NC score. We chose deterministic routing for all multistage conditions because no additional random aspects influenced the routing process. The routing module in the test length condition $I = 15$ and $I = 35$ contained five items. The routing model in the condition with $I = 60$ contained ten items. The cutoff values for the routing into module $\mathbf{m}^{[1]}$ within the first two conditions were $j \leq 2$ and for the third condition $j \leq 5$. Item parameters of all models were drawn from a uniform distribution $\mathrm{U}(-2, 2)$, whereby the item parameters for the routing module $\mathbf{m}^{[2]}$ were from $\mathrm{U}(-1, 1)$, m_1 from $\mathrm{U}(-2, 1)$, and $\mathbf{m}^{[3]}$ from $\mathrm{U}(1, 2)$. In the simulation, four different types of (standardized) distribution of $g(\boldsymbol{\beta})$ were considered (see Figure 2; *skew* as skewness and *kurt* as the kurtosis parameter):

(a) (standard) normal ($skew = 0,\ kurt = 0$): $\theta \sim \mathrm{N}(0, 1)$;
(b) bimodal ($skew = 0.3,\ kurt = -1.0$): $\theta \sim \frac{3}{5}\mathrm{N}(-0.705, 0.254) + \frac{2}{5}\mathrm{N}(1.058, 0.254)$;
(c) skewed ($skew = -1.5,\ kurt = 3.2$): $\theta \sim \frac{1}{5}\mathrm{N}(-1.259, 1.791) + \frac{4}{5}\mathrm{N}(0.315, 0.307)$;
(d) χ_1^2 ($skew = 2.8,\ kurt = 12$): $\theta \sim \chi_1^2$ with one degree of freedom.

The skewed and bimodal distribution parameters were chosen following Casabianca and Lewis [115]. This study also dealt with parameter recovery for MML with log-linear smoothing, but solely in LFT designs. The authors reported that they chose theses parameter based on their own work [127], as well as other contributions that also dealt with simulation studies on the same or related topics (see, e.g., [128–133]).

In disciplines such as educational measurement, clinical psychology, or medicine, there are many situations where the resulting trait distribution might deviate from an assumed normal distribution (see, e.g., [115,129,132,134,135]). A bimodal trait distribution might occur, e.g., in clinical and personality psychology, if one aspect of personality or psychopathology is low for most people and a few people high. One such reported dimension is, e.g., psychoticism, which tends to be positively skewed towards low scores [136]. Furthermore, in situations where groups of persons are examined, in which a subgroup has psychopathological symptoms, distributions deviating from a normal distribution are expected and typically positively skewed [137]. Areas of (large-scale) educational testing, as well as raw scores of state-wide tests tend to be non-normal distributed [138,139].

A bimodal distribution can be expected when two different groups of examinees are investigated, e.g., high versus low performer or schools with privileged versus underprivileged students [140].

For the estimation, the following three different estimation approaches were used:

CML/CMLMST: CML estimation with consideration of the respective MST design in the MST condition (CMLMST);
MMLN: MML estimation, assuming that traits are normally distributed;
MMLS: MML estimation with log-linear smoothing up to four moments.

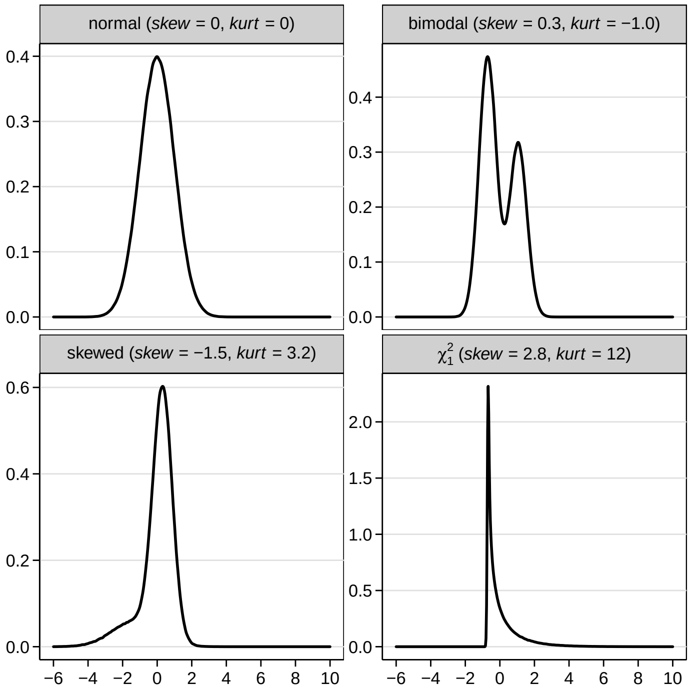

Figure 2. Illustration of the latent trait distributions for all conditions. Note: normal = (standard) normal ability distribution; bimodal: $\theta \sim \frac{3}{5}\mathrm{N}(-0.705, 0.254) + \frac{2}{5}\mathrm{N}(1.058, 0.254)$; skewed: $\theta \sim \frac{1}{5}\mathrm{N}(-1.259, 1.791) + \frac{4}{5}\mathrm{N}(0.315, 0.307)$; χ^2_1: $\theta \sim \chi^2_1$ with one degree of freedom. All distributions are transformed such that $\mathrm{E}(\theta) = 0$ and $\mathrm{Var}(\theta) = 1$.

For each condition, 1000 datasets were generated, and the CML and MML estimation methods were applied. Thereby, 1000 replications *R* were conducted in each cell. For the parameter estimation and the analysis of the simulation study, the open-source software R [124] was used. For reasons of the comparability of the estimated item parameters across the different estimation methods, the estimated item parameters were centered after estimation.

3.2. Implementation in R

All introduced estimation methods were implemented in R packages. For the conventional CML estimation, there is a wide variety of packages available. In addition to the well-known `eRm` with `eRm::RM()` [141], these are, for example, the R packages with the respective functions `psychotools` with `psychotools::raschmodel()`, `immer` with `immer::immer_cml()` and `tmt` with the function `tmt::tmt_rm()`, to name a few representatives [126,141–143]. All packages allow a user-friendly application, but they differ in terms of speed and the availability of further analysis options. With regard to CML parameter estimation in MST designs, two packages are currently available: `dexterMST` with `dexterMST::fit_enorm_mst()` and `tmt` with the function `tmt::tmt_rm()`. The two packages differ concerning the specification of the MST design to be taken into account. In `dexterMST`, first, an MST project must be created with the function `dexterMST::create_mst_project()`, then the the scoring rules used with `dexterMST::add_scoring_rules_mst()` are handed over. Essentially, this is a list of all items, admissible responses, and assigned scores to each response when grading. For the estimation, the routing rules were set with `dexterMST::mst_rules()` and with `dexterMST::create_mst_test()`, then the actual test was carried out, created from the specified rules and the defined modules. Once these steps were executed, the actual data were added with `dexterMST::add_booklet_mst()` to the created database. The actual parameter estimation was realized with `dexterMST::fit_enorm_mst()`. Furthermore, in the `tmt` package, the actual used MST design must be defined. For this purpose, a model language was developed that could be used to define the modules and routing rules. In the first section, the modules were defined, in the example below indicated as m1, m2 and m3. Subsequently, each path of the MST design with the respective rules was specified (in the example below with p1 and p2). In deterministic routing, the lower and upper limit of the raw scores must be specified for each module in each path. The parameter estimation was realized with `tmt::tmt_rm()` with the specified design as an additional argument.

```
model <- ‘‘
 m1 =~c(i01,i02,i03,i04,i05)
 m2 =~c(i06,i07,i08,i09,i10)
 m3 =~paste0(’i’,11:15)

 p1 := m2(0,2) + m1
 p2 := m2(3,5) + m3
’’
```

Furthermore, for MML parameter estimation, numerous packages are available. Some selected examples are `ltm` with `ltm::rasch()`, `sirt` with `sirt::rasch.mml2()` and `TAM` with `TAM::tam.mml()` or `mirt` with `mirt::mirt()`, which also differ in functionality and speed [144–147]. In contrast to CML estimation, no further steps were necessary to obtained the unbiased estimates. The log-linear smoothing used here is available in the package `sirt` [145]. As already pointed out positively by Casabianca and Lewis [115], only the desired number of moments needs to be specified additionally. This can also be emphasized as an advantage compared to the described CML estimation in MST designs, especially in cases with complex MST designs. To utilize the log-linear smoothing, the package `sirt` with the function `sirt::rasch.mirtlc()` is available. The model type (in our case, `modeltype = ”MLC1”`) and the trait distribution `distribution.trait = ”codesmooth4”` were passed as an additional argument (in this example, up to four moments). In the simulation described here, we utilized the R package `sirt` [145] for MML estimation and the R package `tmt` [126] for CML estimation.

3.3. Outcome Measures

To compare the different estimation methods under the different simulation conditions, we computed three criteria. The focus was the estimated item parameters $\hat{\beta}$ in each simulation condition. The computed quantities were the bias of the estimates, the accuracy measured with the root mean squared error (RMSE), and the average relative RMSE (RRMSE) as a summary of the bias and variability. The bias represents the absolute deviation of item parameter estimates from the true item parameter and is reported as the average absolute bias (ABIAS) overall replication in each condition.

$$ABIAS(\hat{\beta}) = \frac{1}{I}\sum_{i=1}^{I}\left|\frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_{ir} - \beta_i\right| = \frac{1}{I}\sum_{i=1}^{I}\left|Bias(\hat{\beta}_i)\right| \tag{17}$$

For the evaluation of the overall accuracy of item parameter estimation, the RMSE was computed. The average RMSE was calculated as the square root of the squared differences between the estimated and true item parameters. The ABIAS and the ARMSE are reported, each as the average for each condition and in the MST case for each module separately.

$$ARMSE(\hat{\beta}) = \frac{1}{I}\sum_{i=1}^{I}\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_{ir} - \beta_i)^2} = \frac{1}{I}\sum_{i=1}^{I}RMSE(\hat{\beta}_i) \tag{18}$$

The RRMSE is defined as follows:

$$RRMSE(\hat{\beta}) = \frac{\sum_{i=1}^{I} RMSE(\hat{\beta}_i)}{\sum_{i=1}^{I} SD_{reference}(\hat{\beta}_i)}, \tag{19}$$

where $SD_{reference}$ is the average standard deviation of the item parameters of the CML method in the fixed-length condition, respectively CMLMST in the MST condition, and serves hereby as the reference.

4. Results

The results of the simulation study are reported separately for the conditions of the LFT and MST. In both conditions, the RMSE is reported in the figures and the ABIAS and RRMSE in the tables. In the simulation, there were no items that all persons did not or wholly solved. Concerning the persons who solved all items or none of the items, the average was 2.5% in the fixed-length condition and 2.6% in the multistage condition. We did not exclude any persons in this regard, but used the default settings of the respective packages. For the item parameter estimation, this was neither a problem for the CML nor the MML estimation method (see, e.g., [148]).

4.1. Results for the Linear Fixed-Length Test Condition

The results for the LFT showed across the estimation conditions very minor differences. Therefore and for a better overview, only the results for the long test condition ($I = 60$) are presented (results for all test lengths and sample sizes can be found in Appendix A Table A1). In Figure 3, the RMSE of all estimation conditions decreased across all trait distribution conditions. There was no difference between the estimation methods either in the normal or in the non-normal conditions (bimodal, skewed, χ^2_1).

The ABIAS and RRMSE reported in Table 1 show very similar results. In the normal distribution condition, there was no difference between the different estimation methods concerning the BIAS of the item parameters. With large sample sizes ($N = 1000$), the MMLS method seemed to lead to a slightly smaller RRMSE compared to CML and MMLN. In the conditions of non-normal distribution, the results were more heterogeneous. In the bimodal condition, the MMLN method with a small sample size ($N = 100$) led to smaller bias, but the difference to CML and MMLS decreased with increasing sample size. The ABIAS in the conditions skewed and χ^2_1 was lower in the CML method, but the difference between CML and MMLS decreased with increasing sample size. It is noteworthy that in the condition skewed, the difference between CML and MMLS was lower than in the condition χ^2_1: here, the CML method led to a smaller bias of the item parameters even with larger sample sizes. Regarding the RRMSE, the MMLS led in both the bimodal, as well as the skewed condition for medium and large sample sizes to the smallest RRMSE. In the χ^2_1 condition, both the CML and MMLS method led to lower RRMSE compared to MMLN. However, it can be summarized that even for the MMLN approach, the results showed compared to the CML and also MMLS condition that the misspecification of the trait distribution had no (large) influence (see also [149]) for a more detailed discussion on different trait distributions in the LFT.

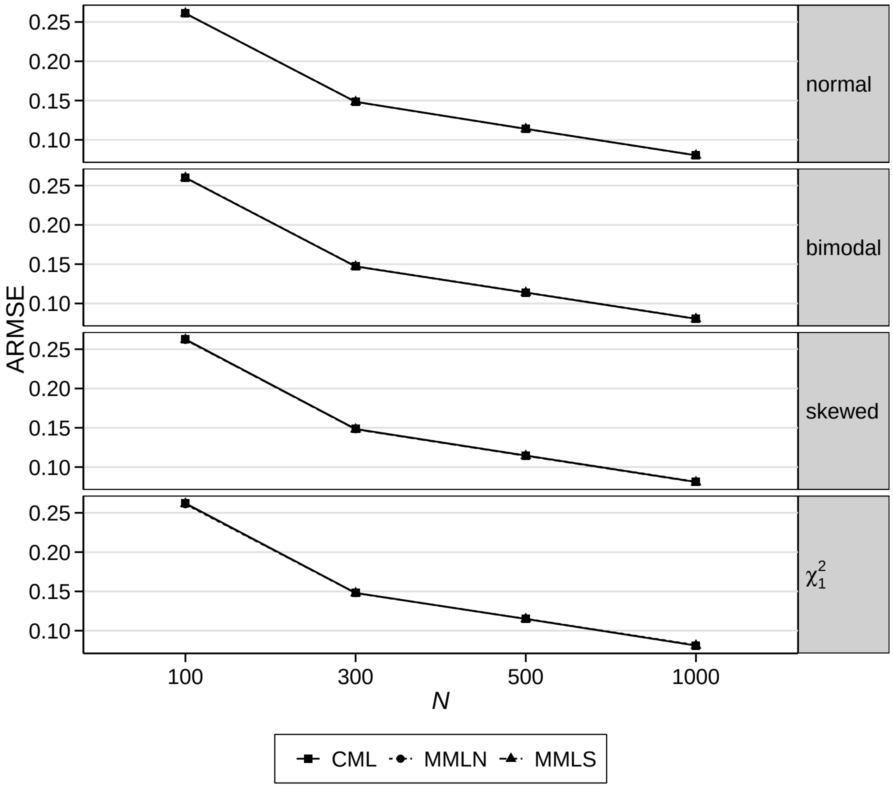

Figure 3. Average root mean squared error (ARMSE) for the fixed-length test condition with 60 items per trait distribution and sample size N. Note: ARMSE = average root mean squared error; normal = (standard) normal ($skew = 0,\ kurt = 0$): $\theta \sim \mathrm{N}(0, 1)$; bimodal ($skew = 0.3,\ kurt = -1.0$): $\theta \sim \frac{3}{5}\mathrm{N}(-0.705, 0.254) + \frac{2}{5}\mathrm{N}(1.058, 0.254)$; skewed ($skew = -1.5,\ kurt = 3.2$): $\theta \sim \frac{1}{5}\mathrm{N}(-1.259, 1.791) + \frac{4}{5}\mathrm{N}(0.315, 0.307)$; χ^2_1 ($skew = 2.8,\ kurt = 12$): $\theta \sim \chi^2_1$ with one degree of freedom. All distributions are transformed such that $\mathrm{E}(\theta) = 0$ and $\mathrm{Var}(\theta) = 1$. CML = conditional maximum likelihood; MMLN = marginal maximum likelihood estimation (MML) with normal distribution; MMLS = MML with log-linear smoothing up to four moments.

4.2. Results for the Multistage Test Condition

The results for the MST condition were more differentiated and therefore discussed separately. For a better overview, the results are not reported separately by module; these can be found in the Appendix A in Figure A1 for the RMSE and two separate tables for the ABIAS in Table A2 and the RRMSE in Table A3. The RMSE in Figure 4 indicates that the conventional CML estimate (i.e., the CML method without considering the respective MST design) led to the largest RMSE across all conditions.

Table 1. Average absolute bias (ABIAS) and relative root mean squared error (RRMSE) for the fixed-length test condition with 60 items as a function of sample size *N* for each trait distribution.

		Normal			**Bimodal**			**Skewed**			χ^2_1		
Criterion	*N*	**CML**	**MMLN**	**MMLS**	**CML**	**MMLN**	**MMLS**	**CML**	**MMLN**	**MMLS**	**CML**	**MMLN**	**MMLS**
ABIAS	100	0.018	0.018	0.018	0.017	0.015	0.018	0.020	0.025	0.021	0.021	0.027	0.022
	300	0.008	0.008	0.008	0.008	0.006	0.008	0.006	0.012	0.007	0.007	0.015	0.009
	500	0.004	0.004	0.004	0.004	0.003	0.004	0.005	0.011	0.005	0.005	0.013	0.007
	1000	0.002	0.002	0.002	0.003	0.003	0.003	0.003	0.009	0.003	0.003	0.012	0.005
RRMSE	100	100.2	100.2	100.3	100.2	100.4	100.3	100.3	100.1	100.2	100.3	99.9	100.1
	300	100.1	100.1	100.1	100.1	100.3	100.1	100.1	100.0	99.8	100.1	100.0	100.0
	500	100.1	100.1	100.0	100.0	100.2	99.9	100.1	100.1	99.8	100.1	100.2	100.1
	1000	100.0	100.0	99.9	100.1	100.2	99.9	100.0	100.4	99.7	100.0	100.8	100.1

Note: ABIAS = average absolute bias; RRMSE = relative root mean squared error with CML as reference; normal = (standard) normal ($skew = 0,\ kurt = 0$): $\theta \sim \mathrm{N}(0,1)$; bimodal ($skew = 0.3,\ kurt = -1.0$): $\theta \sim \frac{3}{5}\mathrm{N}(-0.705, 0.254) + \frac{2}{5}\mathrm{N}(1.058, 0.254)$; skewed ($skew = -1.5,\ kurt = 3.2$): $\theta \sim \frac{1}{5}\mathrm{N}(-1.259, 1.791) + \frac{4}{5}\mathrm{N}(0.315, 0.307)$; χ^2_1 ($skew = 2.8,\ kurt = 12$): $\theta \sim \chi^2_1$ with one degree of freedom. All distributions are transformed such that $\mathrm{E}(\theta) = 0$ and $\mathrm{Var}(\theta) = 1$. CML = conditional maximum likelihood; MMLN = marginal maximum likelihood estimation (MML) with normal distribution; MMLS = MML with log-linear smoothing up to four moments.

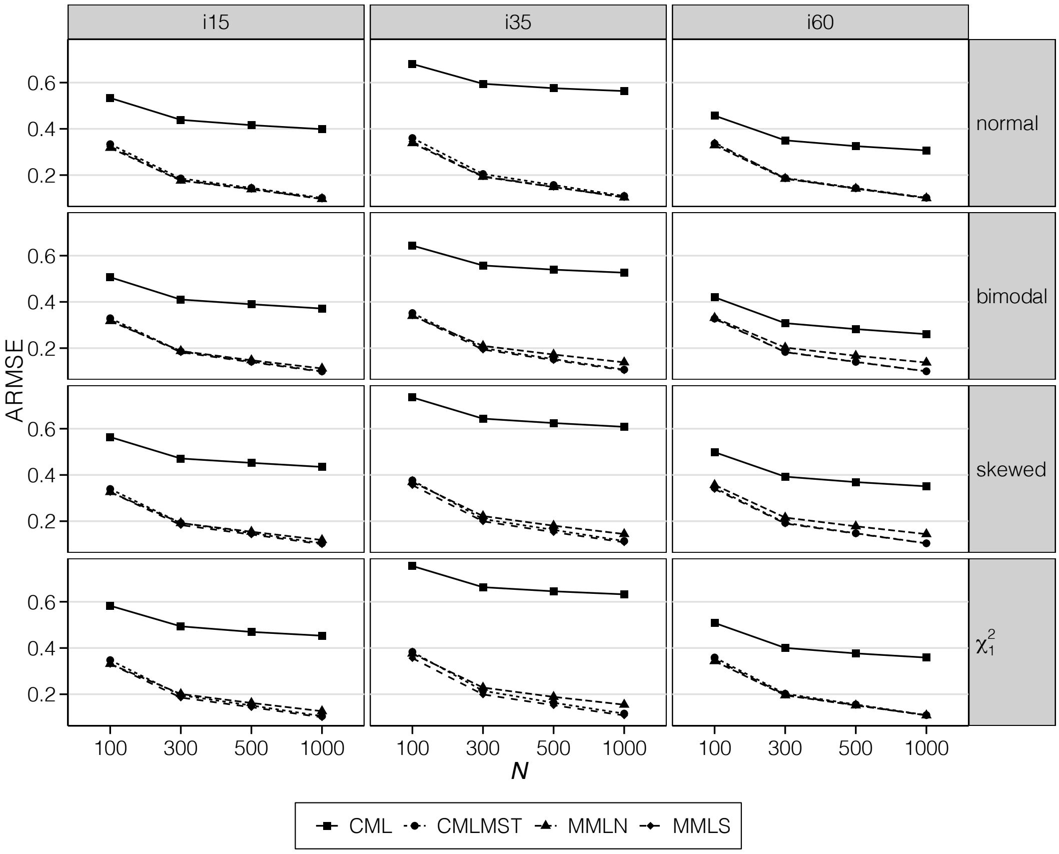

Figure 4. Average root mean squared error (ARMSE) for the multistage condition as a function of sample size N and the number of items for each trait distribution. *Note:* ARMSE = average root mean squared error; normal = (standard) normal (*skew* $= 0$, *kurt* $= 0$): $\theta \sim \text{N}(0, 1)$; bimodal (*skew* $= 0.3$, *kurt* $= -1.0$): $\theta \sim \frac{3}{5}\text{N}(-0.705, 0.254) + \frac{2}{5}\text{N}(1.058, 0.254)$; skewed (*skew* $= -1.5$, *kurt* $= 3.2$): $\theta \sim \frac{1}{5}\text{N}(-1.259, 1.791) + \frac{4}{5}\text{N}(0.315, 0.307)$; χ^2_1 (*skew* $= 2.8$, *kurt* $= 12$): $\theta \sim \chi^2_1$ with one degree of freedom. All distributions are transformed such that $\text{E}(\theta) = 0$ and $\text{Var}(\theta) = 1$. CML = conditional maximum likelihood (CML); CMLMST = CML estimation with consideration of the respective MST design; MMLN = marginal maximum likelihood estimation (MML) with normal distribution; MMLS = MML with log-linear smoothing up to four moments.

4.2.1. Normal Distribution

In Figure 4, the RMSE in the condition with a normal trait distribution was the smallest for the MMLN method. This result was expected because this was the condition with the correct distribution specification. The difference between the estimation methods was small. Concerning the test lengths and sample size, the RMSE of the MMLN method was smaller for short and medium test lengths ($I = 15, 35$) and small sample sizes, but vanished for longer test lengths or sample sizes above $N = 300$. Overall, the difference between the estimation methods in the condition normal distribution except for the CML method seemed to be quite low. With regard to the relative RMSE (RRMSE) in Table 2 at which all estimation methods were referenced to the CMLMST method, these results can be confirmed. Relating to ABIAS, the CMLMST method led to a smaller average bias of the item parameters; however, the difference between CMLMST and MMLN was very small, especially for sample sizes above $N = 100$.

Table 2. Average absolute bias (ABIAS) and relative root mean squared error (RRMSE) for the multistage condition as a function of sample size N and the number of items I for each trait distribution.

			Normal				Bimodal				Skewed				χ^2_1			
Criterion	N	I	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**
ABIAS	100	15	0.023	0.410	0.027	0.033	0.022	0.380	0.030	0.022	0.023	0.447	0.079	0.036	0.029	0.464	0.090	0.029
		35	0.027	0.582	0.026	0.044	0.022	0.544	0.067	0.027	0.034	0.640	0.131	0.060	0.035	0.658	0.140	0.045
		60	0.025	0.314	0.026	0.031	0.027	0.267	0.070	0.025	0.029	0.362	0.126	0.047	0.025	0.368	0.036	0.016
	300	15	0.010	0.389	0.010	0.014	0.005	0.356	0.051	0.006	0.006	0.421	0.064	0.013	0.012	0.445	0.081	0.011
		35	0.007	0.557	0.009	0.017	0.007	0.520	0.087	0.010	0.010	0.608	0.106	0.033	0.009	0.627	0.114	0.021
		60	0.010	0.293	0.009	0.013	0.007	0.245	0.088	0.007	0.008	0.338	0.105	0.025	0.009	0.347	0.026	0.013
	500	15	0.004	0.381	0.004	0.006	0.004	0.355	0.052	0.004	0.005	0.420	0.063	0.014	0.006	0.437	0.077	0.007
		35	0.007	0.552	0.007	0.010	0.004	0.516	0.091	0.007	0.007	0.602	0.100	0.026	0.005	0.623	0.112	0.019
		60	0.006	0.288	0.006	0.008	0.006	0.242	0.090	0.005	0.004	0.334	0.102	0.022	0.005	0.344	0.025	0.016
	1000	15	0.002	0.379	0.003	0.003	0.002	0.351	0.054	0.002	0.003	0.416	0.063	0.011	0.005	0.435	0.076	0.008
		35	0.005	0.551	0.006	0.007	0.004	0.514	0.092	0.007	0.003	0.596	0.097	0.020	0.004	0.620	0.110	0.016
		60	0.005	0.286	0.004	0.005	0.003	0.238	0.094	0.002	0.003	0.332	0.100	0.019	0.004	0.341	0.024	0.017
RRMSE	100	15	100.3	160.4	95.6	97.4	100.2	154.3	96.6	98.3	100.3	166.7	96.4	96.4	100.4	168.3	95.7	95.6
		35	100.3	189.5	94.2	95.9	100.2	183.3	96.7	97.3	100.4	196.1	98.5	94.9	100.5	197.8	98.5	93.3
		60	100.3	136.3	98.1	101.6	100.4	128.6	101.3	99.9	100.4	145.3	103.9	98.8	100.3	142.0	95.9	98.1
	300	15	100.1	236.9	95.0	96.5	100.0	220.8	100.9	98.1	100.0	245.0	99.8	95.4	100.2	252.7	102.9	95.1
		35	100.0	290.7	94.1	95.2	100.0	279.2	105.0	97.0	100.1	307.7	106.2	95.4	100.1	308.5	106.5	92.6
		60	100.1	186.8	98.1	98.9	100.1	167.8	110.3	99.5	100.1	203.8	112.0	98.6	100.1	198.1	96.7	98.1
	500	15	100.0	288.2	96.0	97.0	100.0	274.5	104.0	98.0	100.0	304.7	104.0	96.0	100.1	310.0	107.2	95.0
		35	100.1	366.8	94.4	95.1	100.0	353.6	113.0	97.1	100.1	384.3	111.1	93.9	100.0	395.1	115.5	93.4
		60	100.1	226.2	98.2	98.8	100.1	200.8	118.9	99.5	100.0	249.6	120.4	99.1	100.0	241.9	97.5	98.4
	1000	15	100.0	393.3	95.2	96.4	100.0	372.2	112.4	98.0	100.0	415.7	113.3	95.9	100.1	430.8	120.5	95.4
		35	100.1	515.1	94.4	95.3	100.1	491.5	129.4	97.1	100.0	532.4	126.5	95.2	100.0	543.6	133.2	93.3
		60	100.1	300.0	97.9	98.4	100.0	260.7	137.8	99.6	100.0	336.0	138.0	99.6	100.0	326.7	99.2	99.2

Note: ABIAS = average absolute bias; RRMSE = relative root mean squared error with CMLMST as reference; normal = (standard) normal ($skew = 0,\ kurt = 0$): $\theta \sim \mathrm{N}(0,1)$; bimodal ($skew = 0.3,\ kurt = -1.0$): $\theta \sim \frac{3}{5}\mathrm{N}(-0.705, 0.254) + \frac{2}{5}\mathrm{N}(1.058, 0.254)$; skewed ($skew = -1.5,\ kurt = 3.2$): $\theta \sim \frac{1}{5}\mathrm{N}(-1.259, 1.791) + \frac{4}{5}\mathrm{N}(0.315, 0.307)$; χ^2_1 ($skew = 2.8,\ kurt = 12$): $\theta \sim \chi^2_1$ with one degree of freedom. All distributions are transformed such that $\mathrm{E}(\theta) = 0$ and $\mathrm{Var}(\theta) = 1$. CML = conditional maximum likelihood (CML); CMLMST = CML estimation with consideration of the respective MST design; MMLN = marginal maximum likelihood estimation (MML) with normal distribution; MMLS = MML with log-linear smoothing up to four moments.

4.2.2. Non-Normal Distributions

In the conditions with a non-normal trait distribution, the MMLN method led nearly in all conditions to a higher RRMSE compared to CMLMST and MMLS. Exceptions were the bimodal condition with a small sample size ($N = 100$) together with a short to medium test length ($I = 15, 35$) and the χ^2_1 condition with a long test length ($I = 60$). It should be emphasized that in all other non-normal distribution conditions, the MMLS method led to smaller RRMSE regardless of the sample size and test length compared to MMLN and CMLMST. Concerning the bias of the item parameter in Table 2, the CMLMST method showed the smallest ABIAS independently of sample size and test length. In the bimodal distribution condition, the difference between CMLMST and MMLS was comparatively small, but it should be noted that it was also smaller for the MMLS condition compared to CMLMST. Concerning the two other non-normal distribution conditions (skewed, χ^2_1), the bias of the item parameter in the CMLMST was smaller regardless of sample size and test length.

5. Summary and Discussion

For the estimation of item parameters, alternative estimation methods are available.

While users of the CML method often emphasize that this method comes close to the idea of person-free assessment [148] required for the postulation of specific objectivity [150,151] and that no distribution assumption for the person parameters are required, supporters of the MML method might highlight the flexibility of the approach.

When it comes to MST designs, there was only MML estimation available. If CML parameter estimation were applied, the estimated item parameters would be severely biased. Based on the contribution by Zwitser and Maris [109], two implementations in R packages `dexterMST` [125] and `tmt` [126] are available for item parameter estimation using the CML method in MST designs.

The simulation study was carried out to investigate the influence of trait distributions on the estimation of item parameters. The results showed a differentiated picture. As the sample size increased and the number of items increased, the CMLMST method showed a comparatively small RMSE. As expected, the MMLN method led to a comparatively large RMSE in all non-normal distribution conditions. It is noteworthy that the MMLS estimation method provided the smallest RMSE across conditions. The results were very similar between MMLS and CMLMST, especially with increasing sample sizes and an increasing number of items, even though the MMLS method objectively led to a smaller RMSE. Based on the results, it seems favorable for MST designs to either use the CMLMST or MMLS estimation. Concerning the bias of the item parameter, the CMLMST method led to the smallest ABIAS independently of sample size and test length in nearly all MST conditions. However, in the decision for the CMLMST or MMLS method, it should be considered that the actual distribution used in the MMLS method was assumed to resemble the true population distribution, which may differ. This might be an advantage of the CMLMST method since no distribution assumption was made here.

There are also limitations associated with the present study that might limit the generalizability of the findings. In our research question, we were interested in the influence of the type of trait distribution on item parameter estimation. The number of items and the MST design were varied as additional factors. It would be interesting to systematically study the impact of using more complex MST designs in further studies and perhaps also consider Bayesian estimation methods (see, e.g., [152]). It was noticeable in the results that for the 60-item condition with a χ^2_1 trait distribution, the difference in the RMSE among CML, MMLN, and MMLS was smaller than in the two other item conditions (15 and 35). Next to the different number of items, the MST design in the condition with 60 items differed in the size of the routing module with ten instead of five items. On the other hand, the difference between CML and MMLN seemed to increase with an increasing number of items, but the same size of the routing module. Therefore, it would be interesting to investigate more complex MST designs for item parameter estimation in future research.

Author Contributions: Conceptualization, J.S. and A.R.; methodology, J.S. and A.R.; software, J.S.; validation, J.S. and A.R.; formal analysis, J.S.; investigation, J.S.; writing—original draft preparation, J.S.; writing—review and editing, J.S. and A.R.; visualization, J.S. and A.R.; supervision, A.R.; project administration, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The script for the simulation presented are available on request from the corresponding author.

Acknowledgments: We would like to thank the two anonymous reviewers for their careful reading, comments, and suggestions, which led to an improved final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

1PL	one-parameter logistic model
ABIAS	average absolute bias
ARMSE	average root mean squared error
CAT	computerized adaptive tests
CML	conditional maximum likelihood
CMLMST	CML estimation with consideration of the respective MST design
cNC	cumulative number-correct score
ESF	elementary symmetric function
GDM	general diagnostic model
ILSAs	International Large-Scale Assessments
IRT	item response theory
LFT	linear fixed-length test
LLS	log-linear smoothing
MML	maximum likelihood method
MMLN	MML estimation, assuming that traits are normally distributed
MMLS	MML estimation with log-linear smoothing up to four moments
MST	multistage test
NAEP	National Assessment of Educational Progress
NC	number-correct score
PIAAC	Programme for the International Assessment of Adult Competencies
PISA	Programme for International Student Assessment
RMSE	root mean squared error
RRMSE	relative root mean squared error
TIMSS	Trends in the International Mathematics and Science Study

Appendix A

In Table A1, the results for all test lengths and sample sizes for the linear fixed-length test condition are reported. The results for the multistage condition separately by module and in total can be found in Figure A1 for the RMSE, in Table A2 for the ABIAS, and in Table A3 for the RRMSE.

Table A1. Average absolute bias (ABIAS) and relative root mean squared error (RRMSE) for the fixed-length test condition as a function of sample size N and the number of items I for each trait distribution.

			Normal			**Bimodal**			**Skewed**			χ^2_1		
Criterion	N	I	**CML**	**MMLN**	**MMLS**	**CML**	**MMLN**	**MMLS**	**CML**	**MMLN**	**MMLS**	**CML**	**MMLN**	**MMLS**
ABIAS	100	15	0.017	0.016	0.017	0.021	0.019	0.021	0.020	0.028	0.020	0.020	0.032	0.021
		35	0.017	0.017	0.018	0.018	0.014	0.018	0.019	0.025	0.020	0.022	0.031	0.023
		60	0.018	0.018	0.018	0.017	0.015	0.018	0.020	0.025	0.021	0.021	0.027	0.022
	300	15	0.005	0.005	0.005	0.005	0.006	0.005	0.007	0.019	0.007	0.006	0.024	0.006
		35	0.008	0.008	0.008	0.008	0.006	0.008	0.007	0.015	0.007	0.008	0.019	0.010
		60	0.008	0.008	0.008	0.008	0.006	0.008	0.006	0.012	0.007	0.007	0.015	0.009
	500	15	0.003	0.003	0.003	0.004	0.007	0.004	0.003	0.019	0.003	0.004	0.024	0.005
		35	0.005	0.005	0.004	0.005	0.004	0.005	0.005	0.013	0.005	0.004	0.016	0.006
		60	0.004	0.004	0.004	0.004	0.003	0.004	0.005	0.011	0.005	0.005	0.013	0.007
	1000	15	0.002	0.002	0.002	0.003	0.008	0.002	0.002	0.018	0.003	0.003	0.024	0.003
		35	0.002	0.002	0.002	0.003	0.005	0.003	0.003	0.012	0.003	0.003	0.016	0.005
		60	0.002	0.002	0.002	0.003	0.003	0.003	0.003	0.009	0.003	0.003	0.012	0.005
RRMSE	100	15	100.2	100.1	100.2	100.4	100.5	100.5	100.4	99.8	100.1	100.4	99.6	100.1
		35	100.2	100.2	100.2	100.3	100.4	100.3	100.3	99.9	100.0	100.4	99.8	100.1
		60	100.2	100.2	100.3	100.2	100.4	100.3	100.3	100.1	100.2	100.3	99.9	100.1
	300	15	100.0	99.9	99.9	100.1	100.3	100.0	100.1	100.5	99.7	100.1	100.8	99.8
		35	100.1	100.1	100.0	100.1	100.3	100.0	100.1	100.1	99.7	100.1	100.3	100.0
		60	100.1	100.1	100.1	100.1	100.3	100.1	100.1	100.0	99.8	100.1	100.0	100.0
	500	15	100.0	100.0	99.9	100.0	100.5	99.9	100.0	101.2	99.6	100.1	102.0	99.9
		35	100.1	100.0	99.9	100.1	100.3	99.9	100.1	100.4	99.6	100.0	100.5	99.9
		60	100.1	100.1	100.0	100.0	100.2	99.9	100.1	100.1	99.8	100.1	100.2	100.1
	1000	15	100.0	100.0	99.8	100.0	100.8	99.8	100.0	102.7	99.6	100.0	104.8	100.0
		35	100.0	100.0	99.8	100.0	100.5	99.8	100.0	100.9	99.6	100.1	101.9	100.0
		60	100.0	100.0	99.9	100.1	100.2	99.9	100.0	100.4	99.7	100.0	100.8	100.1

Note: ABIAS = average absolute bias; RRMSE = relative root mean squared error with CML as reference; normal = (standard) normal ($skew = 0,\ kurt = 0$): $\theta \sim \mathrm{N}(0,1)$; bimodal ($skew = 0.3,\ kurt = -1.0$): $\theta \sim \frac{3}{5}\mathrm{N}(-0.705, 0.254) + \frac{2}{5}\mathrm{N}(1.058, 0.254)$; skewed ($skew = -1.5,\ kurt = 3.2$): $\theta \sim \frac{1}{5}\mathrm{N}(-1.259, 1.791) + \frac{4}{5}\mathrm{N}(0.315, 0.307)$; χ^2_1 ($skew = 2.8,\ kurt = 12$): $\theta \sim \chi^2_1$ with one degree of freedom. All distributions are transformed such that $\mathrm{E}(\theta) = 0$ and $\mathrm{Var}(\theta) = 1$. CML = conditional maximum likelihood; MMLN = marginal maximum likelihood estimation (MML) with normal distribution; MMLS = MML with log-linear smoothing up to four moments.

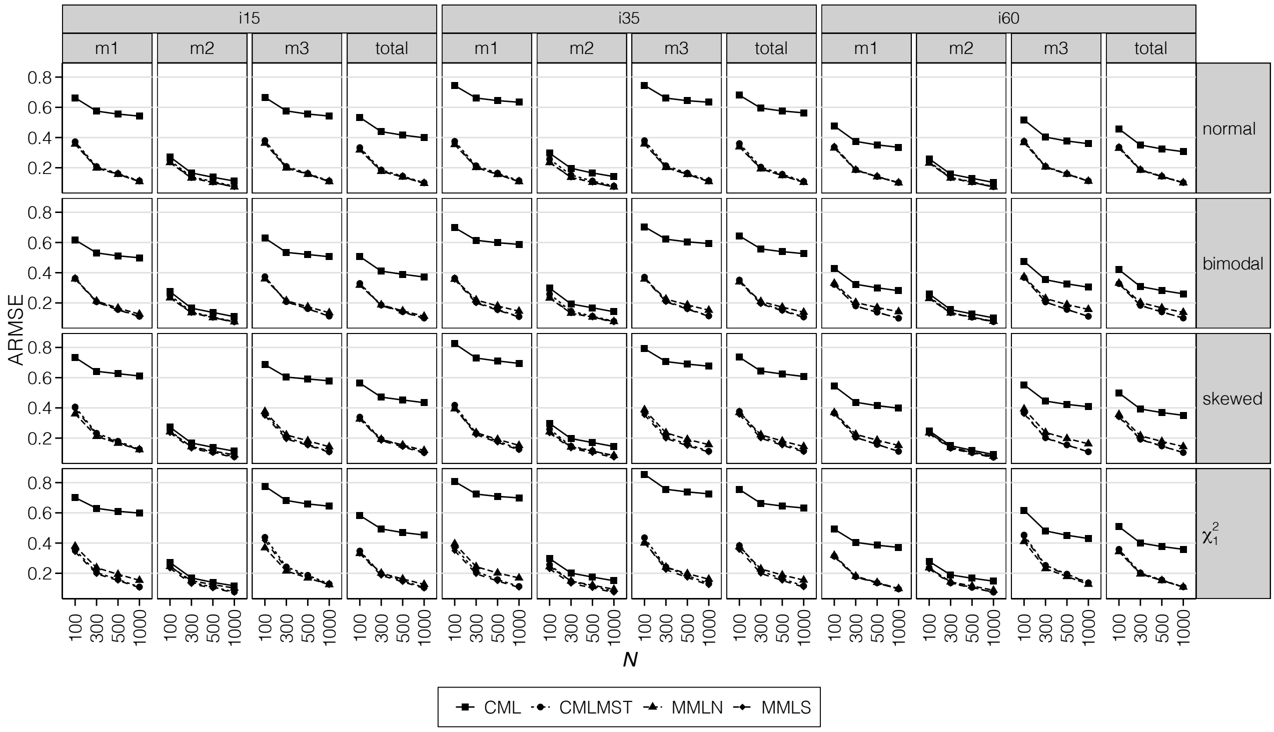

Figure A1. Average root mean squared error (ARMSE) for the multistage condition as a function of sample size N and the number of items for each module separately and in total for each trait distribution. Note: ARMSE = average root mean squared error; normal = (standard) normal (*skew* $= 0$, *kurt* $= 0$): $\theta \sim \mathrm{N}(0, 1)$; bimodal (*skew* $= 0.3$, *kurt* $= -1.0$): $\theta \sim \frac{3}{5}\mathrm{N}(-0.705, 0.254) + \frac{2}{5}\mathrm{N}(1.058, 0.254)$; skewed (*skew* $= -1.5$, *kurt* $= 3.2$): $\theta \sim \frac{1}{5}\mathrm{N}(-1.259, 1.791) + \frac{4}{5}\mathrm{N}(0.315, 0.307)$; χ^2_1 (*skew* $= 2.8$, *kurt* $= 12$): $\theta \sim \chi^2_1$ with one degree of freedom. All distributions are transformed such that $\mathrm{E}(\theta) = 0$ and $\mathrm{Var}(\theta) = 1$. CML = conditional maximum likelihood (CML); CMLMST = CML estimation with consideration of the respective MST design; MMLN = marginal maximum likelihood estimation (MML) with normal distribution; MMLS = MML with log-linear smoothing up to four moments.

Table A2. Average absolute bias (ABIAS) for the multistage condition as a function of sample size *N* and the number of items *I* for each module separately and in total for each trait distribution.

			Normal				Bimodal				Skewed				χ^2_1			
N	*I*	**Modules**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**
100	15	m1	0.031	0.574	0.037	0.044	0.027	0.525	0.031	0.025	0.028	0.641	0.070	0.048	0.034	0.625	0.135	0.042
		m2	0.010	0.081	0.010	0.011	0.010	0.081	0.015	0.009	0.016	0.095	0.051	0.019	0.013	0.091	0.072	0.008
		m3	0.028	0.574	0.035	0.044	0.030	0.536	0.045	0.031	0.026	0.606	0.116	0.041	0.040	0.674	0.064	0.038
		total	0.023	0.410	0.027	0.033	0.022	0.380	0.030	0.022	0.023	0.447	0.079	0.036	0.029	0.464	0.090	0.029
	35	m1	0.030	0.660	0.028	0.049	0.025	0.613	0.070	0.031	0.036	0.733	0.138	0.068	0.037	0.733	0.162	0.051
		m2	0.013	0.118	0.012	0.015	0.011	0.118	0.024	0.010	0.016	0.130	0.050	0.019	0.015	0.129	0.064	0.021
		m3	0.030	0.659	0.028	0.048	0.022	0.617	0.078	0.029	0.037	0.716	0.150	0.065	0.038	0.758	0.143	0.047
		total	0.027	0.582	0.026	0.044	0.022	0.544	0.067	0.027	0.034	0.640	0.131	0.060	0.035	0.658	0.140	0.045
	60	m1	0.027	0.351	0.028	0.034	0.029	0.296	0.074	0.026	0.034	0.419	0.138	0.055	0.027	0.389	0.042	0.018
		m2	0.011	0.070	0.012	0.012	0.015	0.066	0.024	0.014	0.009	0.054	0.035	0.011	0.011	0.132	0.047	0.015
		m3	0.028	0.374	0.030	0.036	0.030	0.319	0.083	0.028	0.033	0.428	0.150	0.053	0.029	0.442	0.026	0.015
		total	0.025	0.314	0.026	0.031	0.027	0.267	0.070	0.025	0.029	0.362	0.126	0.047	0.025	0.368	0.036	0.016
300	15	m1	0.012	0.545	0.011	0.017	0.007	0.497	0.051	0.008	0.008	0.609	0.049	0.018	0.014	0.603	0.119	0.017
		m2	0.005	0.075	0.005	0.006	0.003	0.070	0.024	0.004	0.003	0.080	0.048	0.005	0.005	0.084	0.070	0.008
		m3	0.014	0.546	0.013	0.019	0.005	0.501	0.076	0.004	0.007	0.576	0.095	0.015	0.016	0.648	0.055	0.009
		total	0.010	0.389	0.010	0.014	0.005	0.356	0.051	0.006	0.006	0.421	0.064	0.013	0.012	0.445	0.081	0.011
	35	m1	0.008	0.632	0.010	0.019	0.006	0.585	0.094	0.010	0.011	0.698	0.112	0.038	0.008	0.698	0.134	0.025
		m2	0.004	0.108	0.004	0.005	0.004	0.111	0.024	0.005	0.005	0.118	0.035	0.012	0.008	0.121	0.054	0.013
		m3	0.007	0.631	0.009	0.018	0.008	0.590	0.101	0.012	0.011	0.680	0.123	0.035	0.009	0.724	0.115	0.021
		total	0.007	0.557	0.009	0.017	0.007	0.520	0.087	0.010	0.010	0.608	0.106	0.033	0.009	0.627	0.114	0.021
	60	m1	0.011	0.329	0.011	0.015	0.009	0.273	0.094	0.008	0.008	0.391	0.113	0.029	0.010	0.367	0.023	0.008
		m2	0.004	0.061	0.003	0.004	0.003	0.057	0.029	0.004	0.005	0.052	0.033	0.007	0.005	0.124	0.050	0.021
		m3	0.011	0.350	0.011	0.015	0.008	0.293	0.105	0.007	0.008	0.401	0.126	0.028	0.009	0.416	0.020	0.015
		total	0.010	0.293	0.009	0.013	0.007	0.245	0.088	0.007	0.008	0.338	0.105	0.025	0.009	0.347	0.026	0.013

Table A2. *Cont.*

			Normal				Bimodal				Skewed				χ^2_1			
N	*I*	**Modules**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**
500	15	m1	0.005	0.536	0.005	0.008	0.003	0.492	0.057	0.004	0.008	0.606	0.048	0.020	0.008	0.593	0.110	0.008
		m2	0.002	0.071	0.001	0.002	0.004	0.074	0.023	0.004	0.004	0.080	0.047	0.007	0.002	0.078	0.070	0.009
		m3	0.006	0.536	0.005	0.009	0.005	0.500	0.077	0.005	0.003	0.574	0.095	0.016	0.009	0.638	0.051	0.004
		total	0.004	0.381	0.004	0.006	0.004	0.355	0.052	0.004	0.005	0.420	0.063	0.014	0.006	0.437	0.077	0.007
	35	m1	0.007	0.627	0.008	0.011	0.004	0.581	0.098	0.008	0.008	0.691	0.104	0.029	0.005	0.694	0.131	0.022
		m2	0.004	0.105	0.002	0.003	0.004	0.111	0.026	0.005	0.004	0.118	0.039	0.008	0.002	0.120	0.054	0.013
		m3	0.007	0.626	0.007	0.011	0.004	0.585	0.105	0.008	0.007	0.674	0.117	0.028	0.006	0.719	0.113	0.018
		total	0.007	0.552	0.007	0.010	0.004	0.516	0.091	0.007	0.007	0.602	0.100	0.026	0.005	0.623	0.112	0.019
	60	m1	0.006	0.323	0.006	0.009	0.006	0.269	0.097	0.005	0.005	0.386	0.111	0.026	0.005	0.363	0.019	0.011
		m2	0.004	0.061	0.004	0.004	0.003	0.058	0.027	0.003	0.003	0.047	0.030	0.005	0.003	0.124	0.050	0.020
		m3	0.007	0.344	0.006	0.009	0.007	0.289	0.108	0.006	0.004	0.396	0.123	0.025	0.006	0.412	0.020	0.019
		total	0.006	0.288	0.006	0.008	0.006	0.242	0.090	0.005	0.004	0.334	0.102	0.022	0.005	0.344	0.025	0.016
1000	15	m1	0.003	0.532	0.003	0.004	0.003	0.489	0.059	0.003	0.003	0.601	0.046	0.015	0.005	0.590	0.107	0.008
		m2	0.001	0.071	0.001	0.001	0.002	0.070	0.023	0.002	0.002	0.077	0.049	0.004	0.003	0.081	0.069	0.010
		m3	0.002	0.533	0.003	0.004	0.001	0.496	0.080	0.001	0.004	0.571	0.093	0.014	0.006	0.634	0.051	0.006
		total	0.002	0.379	0.003	0.003	0.002	0.351	0.054	0.002	0.003	0.416	0.063	0.011	0.005	0.435	0.076	0.008
	35	m1	0.006	0.625	0.006	0.008	0.004	0.579	0.099	0.007	0.004	0.685	0.100	0.023	0.005	0.691	0.129	0.019
		m2	0.002	0.106	0.002	0.002	0.002	0.109	0.026	0.003	0.003	0.114	0.038	0.005	0.002	0.118	0.054	0.014
		m3	0.006	0.625	0.007	0.008	0.005	0.584	0.106	0.008	0.003	0.668	0.113	0.022	0.005	0.716	0.110	0.014
		total	0.005	0.551	0.006	0.007	0.004	0.514	0.092	0.007	0.003	0.596	0.097	0.020	0.004	0.620	0.110	0.016
	60	m1	0.005	0.321	0.004	0.005	0.003	0.264	0.102	0.003	0.003	0.383	0.108	0.022	0.004	0.360	0.017	0.013
		m2	0.003	0.061	0.002	0.002	0.002	0.057	0.027	0.001	0.002	0.048	0.031	0.004	0.002	0.125	0.048	0.019
		m3	0.006	0.342	0.005	0.006	0.003	0.285	0.112	0.002	0.003	0.394	0.121	0.022	0.004	0.410	0.020	0.021
		total	0.005	0.286	0.004	0.005	0.003	0.238	0.094	0.002	0.003	0.332	0.100	0.019	0.004	0.341	0.024	0.017

Note: ABIAS = average absolute bias; normal = (standard) normal (*skew* $= 0$, *kurt* $= 0$): $\theta \sim \mathrm{N}(0,1)$; bimodal (*skew* $= 0.3$, *kurt* $= -1.0$): $\theta \sim \frac{3}{5}\mathrm{N}(-0.705, 0.254) + \frac{2}{5}\mathrm{N}(1.058, 0.254)$; skewed (*skew* $= -1.5$, *kurt* $= 3.2$): $\theta \sim \frac{1}{5}\mathrm{N}(-1.259, 1.791) + \frac{4}{5}\mathrm{N}(0.315, 0.307)$; χ^2_1 (*skew* $= 2.8$, *kurt* $= 12$): $\theta \sim \chi^2_1$ with one degree of freedom. All distributions are transformed such that $\mathrm{E}(\theta) = 0$ and $\mathrm{Var}(\theta) = 1$. CML = conditional maximum likelihood (CML); CMLMST = CML estimation with consideration of the respective MST design; MMLN = marginal maximum likelihood estimation (MML) with normal distribution; MMLS = MML with log-linear smoothing up to four moments.

Table A3. Relative root mean squared error (RRMSE) for the multistage condition as a function of sample size *N* and the number of items *I* for each module separately and in total for each trait distribution.

			Normal				Bimodal				Skewed				χ^2_1			
N	***I***	**Modules**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**
100	15	M1	100.4	178.6	96.1	97.9	100.2	170.2	98.9	98.7	100.3	181.1	89.0	97.3	100.5	198.0	107.2	96.6
		M2	100.1	109.6	94.5	96.0	100.1	109.6	93.4	96.0	100.2	109.9	97.5	94.9	100.1	109.2	98.8	93.9
		M3	100.3	175.7	95.7	97.9	100.3	168.8	96.4	99.3	100.3	189.5	103.9	96.3	100.5	177.9	84.6	95.9
		total	100.3	160.4	95.6	97.4	100.2	154.3	96.6	98.3	100.3	166.7	96.4	96.4	100.4	168.3	95.7	95.6
	35	M1	100.3	199.0	94.6	96.5	100.2	192.2	98.2	97.6	100.4	198.2	94.6	95.9	100.5	217.6	106.6	94.2
		M2	100.2	115.7	91.4	92.2	100.1	115.6	89.9	92.4	100.3	115.0	93.4	90.2	100.2	116.1	95.2	89.3
		M3	100.3	196.7	94.4	96.1	100.2	190.3	96.8	98.2	100.5	212.6	104.1	95.0	100.4	197.0	92.3	93.4
		total	100.3	189.5	94.2	95.9	100.2	183.3	96.7	97.3	100.4	196.1	98.5	94.9	100.5	197.8	98.5	93.3
	60	M1	100.3	141.9	98.5	102.5	100.4	133.6	102.8	100.0	100.4	149.5	101.4	99.4	100.4	157.6	102.2	99.0
		M2	100.1	108.3	96.4	98.1	100.2	108.4	95.7	98.0	100.1	105.3	99.5	96.9	100.1	117.6	99.9	96.6
		M3	100.3	138.4	98.2	101.7	100.4	129.5	101.5	100.2	100.4	151.4	107.5	98.7	100.2	136.4	90.7	97.8
		total	100.3	136.3	98.1	101.6	100.4	128.6	101.3	99.9	100.4	145.3	103.9	98.8	100.3	142.0	95.9	98.1
300	15	M1	100.2	276.8	95.4	97.1	100.1	255.8	102.3	98.8	100.0	276.2	90.9	96.4	100.3	308.2	115.9	96.1
		M2	100.0	118.4	94.1	95.0	100.0	116.9	95.7	95.7	100.0	119.6	102.1	94.1	100.0	121.2	108.1	93.5
		M3	100.2	276.9	95.3	96.9	100.0	256.3	103.1	99.1	100.0	295.3	108.3	95.0	100.2	281.7	89.0	95.1
		total	100.1	236.9	95.0	96.5	100.0	220.8	100.9	98.1	100.0	245.0	99.8	95.4	100.2	252.7	102.9	95.1
	35	M1	100.0	310.2	94.6	95.8	100.0	299.8	107.1	97.4	100.1	315.4	102.0	96.2	100.1	343.4	115.7	93.2
		M2	100.0	129.2	89.7	90.1	100.0	133.0	90.9	91.9	100.0	134.3	96.1	90.9	100.1	137.7	101.7	90.6
		M3	100.0	309.6	94.6	95.8	100.1	292.8	106.2	97.8	100.1	340.0	113.3	95.4	100.1	312.5	99.4	92.5
		total	100.0	290.7	94.1	95.2	100.0	279.2	105.0	97.0	100.1	307.7	106.2	95.4	100.1	308.5	106.5	92.6
	60	M1	100.2	200.3	98.5	99.1	100.1	179.9	113.0	99.7	100.0	212.4	109.3	99.2	100.1	227.7	101.7	98.6
		M2	100.0	116.0	96.8	97.2	100.0	114.1	98.1	97.9	100.0	111.1	102.5	97.1	100.1	138.7	104.5	97.4
		M3	100.1	193.1	98.1	99.2	100.0	171.4	111.2	99.8	100.1	219.9	117.2	98.5	100.0	190.1	91.4	97.8
		total	100.1	186.8	98.1	98.9	100.1	167.8	110.3	99.5	100.1	203.8	112.0	98.6	100.1	198.1	96.7	98.1

Table A3. *Cont.*

			Normal				Bimodal				Skewed				χ^2_1			
N	***I***	**Modules**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**	**CMLMST**	**CML**	**MMLN**	**MMLS**
500	15	M1	100.1	343.5	96.7	98.0	100.0	326.4	105.9	98.3	100.1	349.7	92.8	96.4	100.1	384.1	121.7	96.0
		M2	100.0	125.8	94.2	94.7	100.1	128.2	95.9	96.1	100.0	129.3	106.4	95.2	100.0	128.6	115.0	93.5
		M3	100.0	342.4	96.5	97.6	100.0	320.9	107.5	99.1	100.0	373.3	115.2	96.1	100.1	352.6	90.4	95.1
		total	100.0	288.2	96.0	97.0	100.0	274.5	104.0	98.0	100.0	304.7	104.0	96.0	100.1	310.0	107.2	95.0
	35	M1	100.1	392.6	94.8	95.5	100.0	383.3	115.2	97.3	100.1	392.8	106.0	95.0	100.0	448.0	127.7	94.2
		M2	100.0	146.8	91.8	91.8	100.0	149.5	94.7	93.7	100.0	147.3	98.6	89.0	100.0	152.2	105.1	90.0
		M3	100.1	391.2	94.7	95.4	100.0	372.1	115.0	97.7	100.1	432.0	120.0	93.9	100.0	400.3	107.2	93.3
		total	100.1	366.8	94.4	95.1	100.0	353.6	113.0	97.1	100.1	384.3	111.1	93.9	100.0	395.1	115.5	93.4
	60	M1	100.1	246.9	98.5	99.0	100.1	216.8	122.6	99.6	100.0	262.6	118.0	99.9	100.1	282.9	102.2	99.0
		M2	100.0	121.5	96.7	97.0	100.0	120.9	99.7	97.9	100.0	114.5	104.1	96.9	100.0	156.6	108.6	98.1
		M3	100.1	235.7	98.3	99.2	100.1	208.2	120.7	99.8	100.0	272.7	127.2	99.0	100.0	232.0	91.8	98.0
		total	100.1	226.2	98.2	98.8	100.1	200.8	118.9	99.5	100.0	249.6	120.4	99.1	100.0	241.9	97.5	98.4
1000	15	M1	100.0	476.3	95.4	96.8	100.0	449.5	113.5	98.8	100.0	483.0	97.4	96.4	100.1	546.2	140.2	96.1
		M2	100.0	144.0	93.8	94.3	100.0	143.9	99.1	95.2	100.0	148.0	116.3	94.5	100.1	152.2	132.6	94.1
		M3	100.0	481.4	96.1	97.5	100.0	449.9	120.4	99.2	100.1	524.6	129.5	96.4	100.0	498.4	96.5	95.5
		total	100.0	393.3	95.2	96.4	100.0	372.2	112.4	98.0	100.0	415.7	113.3	95.9	100.1	430.8	120.5	95.4
	35	M1	100.1	554.3	94.7	95.8	100.1	536.0	132.8	97.5	100.0	549.1	120.1	96.2	100.1	618.9	149.8	94.1
		M2	100.0	178.9	91.7	91.8	100.0	179.9	97.7	92.6	100.0	180.9	106.5	90.4	100.0	187.5	117.5	90.9
		M3	100.1	553.4	94.7	95.8	100.1	520.7	133.3	97.9	100.0	596.3	138.5	95.2	100.0	551.4	122.2	93.0
		total	100.1	515.1	94.4	95.3	100.1	491.5	129.4	97.1	100.0	532.4	126.5	95.2	100.0	543.6	133.2	93.3
	60	M1	100.1	328.6	98.0	98.4	100.0	285.1	144.1	99.7	100.0	355.6	135.2	100.4	100.1	386.6	103.0	99.5
		M2	100.1	138.5	97.1	97.3	100.0	134.0	103.9	97.8	100.0	125.2	109.5	96.2	100.0	198.1	117.4	99.3
		M3	100.1	316.8	98.0	98.6	100.0	273.6	141.5	99.9	100.0	372.5	148.5	99.7	100.0	312.8	92.6	99.0
		total	100.1	300.0	97.9	98.4	100.0	260.7	137.8	99.6	100.0	336.0	138.0	99.6	100.0	326.7	99.2	99.2

Note: RRMSE = relative root mean squared error with CMLMST as reference; normal = (standard) normal ($skew = 0,\ kurt = 0$): $\theta \sim \mathrm{N}(0,1)$; bimodal ($skew = 0.3,\ kurt = -1.0$): $\theta \sim \frac{3}{5}\mathrm{N}(-0.705, 0.254) + \frac{2}{5}\mathrm{N}(1.058, 0.254)$; skewed ($skew = -1.5,\ kurt = 3.2$): $\theta \sim \frac{1}{5}\mathrm{N}(-1.259, 1.791) + \frac{4}{5}\mathrm{N}(0.315, 0.307)$; χ^2_1 ($skew = 2.8,\ kurt = 12$): $\theta \sim \chi^2_1$ with one degree of freedom. All distributions are transformed such that $\mathrm{E}(\theta) = 0$ and $\mathrm{Var}(\theta) = 1$. CML = conditional maximum likelihood (CML); CMLMST = CML estimation with consideration of the respective MST design; MMLN = marginal maximum likelihood estimation (MML) with normal distribution; MMLS = MML with log-linear smoothing up to four moments.

References

1. Lord, F.M. Robbins-Monro procedures for tailored testing. *Educ. Psychol. Meas.* **1971**, *31*, 3–31. [CrossRef]
2. Owen, R.J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *J. Am. Stat. Assoc.* **1975**, *70*, 351–356. [CrossRef]
3. Weiss, D.J. Adaptive Testing Research in Minnesota: Overview, Recent Results, and Future Directions. In *Proceedings of the First Conference on Computerized Adaptive Testing*; Clark, C.L., Ed.; US Civil Service Commission, Personnel Research and Development Center: Washington, DC, USA, 1976; Volume 75, pp. 24–35.
4. Weiss, D.J. New Horizons in Testing. In *Latent Trait Test Theory and Computerized Adaptive Testing*; Academic Press: New York, NY, USA, 1983. [CrossRef]
5. Van der Linden, W.J.; Glas, C.A. *Elements of Adaptive Testing*; Springer: New York, NY, USA, 2010. [CrossRef]
6. Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*; Hillsdale, N.J., Ed.; Erlbaum: Mahwah, Bergen, 1980. [CrossRef]
7. Wainer, H.; Dorans, N.J.; Flaugher, R.; Green, B.F.; Mislevy, R.J.; Steinberg, L.; Thissen, D. *Computerized Adaptive Testing: A Primer*, 2nd ed.; Lawrence Erlbaum: Hillsdale, NJ, USA, 2000.
8. Angoff, W.; Huddleston, E. *The Multi-Level Experiment: A Study of a Two-Level Test System for the College Board Scholastic Aptitude Test*; Statistical Report SR-58-21; Educational Testing Service: Princeton, NJ, USA, 1958.
9. Lord, F.M.; Novick, M.R.; Birnbaum, A. *Statistical Theories of Mental Test Scores*; Addison-Wesley: Menlo Park, CA, USA, 1968.
10. Lord, F.M. Some test theory for tailored testing. *ETS Research Bulletin Series* **1968**, *1968*, i-62. [CrossRef]
11. Lord, F.M. A theoretical study of two-stage testing. *Psychometrika* **1971**, *36*, 227–242. [CrossRef]
12. Zenisky, A.; Hambleton, R.K.; Luecht, R.M. Multistage testing: Issues, designs, and research. In *Elements of Adaptive Testing*; van der Linden, W.J., Glas, C.A., Eds.; Springer: New York, NY, USA, 2009; pp. 355–372. [CrossRef]
13. Luecht, R.M.; Nungester, R.J. Some practical examples of computer-adaptive sequential testing. *J. Educ. Meas.* **1998**, *35*, 229–249. [CrossRef]
14. Weiss, D.J. Improving measurement quality and efficiency with adaptive testing. *Appl. Psychol. Meas.* **1982**, *6*, 473–492. [CrossRef]
15. Hendrickson, A. An NCME instructional module on multistage testing. *Educ. Meas. Issues Pract.* **2007**, *26*, 44–52. [CrossRef]
16. Chang, H.H. Psychometrics behind computerized adaptive testing. *Psychometrika* **2015**, *80*, 1–20. [CrossRef] [PubMed]
17. Betz, N.E.; Weiss, D.J. *Simulation Studies of Two-Stage Ability Testing*; Research Report 74-4; Psychometric Methods Program, Department of Psychology, University of Minnesota: Minneapolis, MN, USA, 1974.
18. Kim, H.; Plake, B.S. Monte Carlo Simulation Comparison of Two-Stage Testing and Computerized Adaptive Testing. Doctoral Dissertation, The University of Nebraska-Lincoln, Lincoln, NE, USA, April 1993.
19. Linn, R.L.; Rock, D.A.; Cleary, T.A. The development and evaluation of several programmed testing methods. *Educ. Psychol. Meas.* **1969**, *29*, 129–146. [CrossRef]
20. Jodoin, M.G.; Zenisky, A.; Hambleton, R.K. Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Appl. Meas. Educ.* **2006**, *19*, 203–220._3. [CrossRef]
21. Weiss, D.J.; Kingsbury, G.G. Application of computerized adaptive testing to educational problems. *J. Educ. Meas.* **1984**, *21*, 361–375. [CrossRef]
22. Cronbach, L.J.; Gleser, G.C. *Psychological Tests and Personnel Decisions*; University of Illinois Press: Urbana, IL, USA, 1957.
23. Schnipke, D.L.; Reese, L.M. *A Comparison of Testlet-Based Test Designs for Computerized Adaptive Testing*; American Educational Research Asociation: Chicago, IL, USA, March 1997.
24. Lord, F.M. Practical methods for redesigning a homogeneous test, also for designing a multilevel test. *ETS Res. Bull. Ser.* **1974**, *1974*, i-26. [CrossRef]
25. Organisation for Economic Co-operation and Development. *PISA 2018 Assessment and Analytical Framework*; OECD Publishing: Paris, France, 2019. [CrossRef]
26. Organisation for Economic Co-operation and Development. *Technical Report of the Survey of Adult Skills (PIAAC)*, 3rd ed.; OECD Publishing: Paris, France, 2019.
27. Fishbein, B.; Martin, M.O.; Mullis, I.V.; Foy, P. The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-Scale Assess. Educ.* **2018**, *6*, 1–23. [CrossRef]
28. Campbell, J.R.; Hombo, C.M.; Mazzeo, J. *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance*; NCES 2000-469; National Center for Educational Statistic: Washington, DC, USA, 2000.
29. Zhang, T.; Xie, Q.; Park, B.J.; Kim, Y.Y.; Broer, M.; Bohrnstedt, G. *Computer Familiarity and Its Relationship to Performance in Three NAEP Digital-Based Assessments*; AIR-NAEP Working Paper #01-2016; American Institutes for Research: Washington, DC, USA, 2016.
30. Kubinger, K.; Holocher-Ertl, S. *AID 3: Adaptives Intelligenz Diagnostikum 3 [AID 3: Adaptive Intelligence Diagnostic 3]*; Beltz-Test: Göttingen, Germany, 2014.
31. Dean, V.; Martineau, J. A state perspective on enhancing assessment and accountability systems through systematic implementation of technology. In *Computers and their Impact on State Assessment: Recent History and Predictions for the Future*; Lissitz, R.W., Jiao, H., Eds.; Information Age Publishing, Inc.: Charlotte, NC, USA, 2012; pp. 25–53.

32. Chen, H.; Yamamoto, K.; von Davier, M. Controlling multistage testing exposure rates in international large-scale assessments. In *Computerized Multistage Testing: Theory and Applications*; Yan, D., von Davier, A.A., Lewis, C., Eds.; CRC Press: New York, NY, USA, 2014; pp. 391–409. [CrossRef]
33. Han, K.C.T.; Guo, F. Multistage testing by shaping modules on the fly. In *Computerized Multistage Testing: Theory and Applications*; Yan, A., von Davier, A.A., Lewis, C., Eds.; CRC Press: New York, NY, USA, 2014; pp. 119–133. [CrossRef]
34. Zheng, Y.; Chang, H.H. On-the-fly assembled multistage adaptive testing. *Appl. Psychol. Meas.* **2014**, *39*, 104–118. [CrossRef]
35. Luo, X.; Wang, X. Dynamic multistage testing: A highly efficient and regulated adaptive testing method. *Int. J. Test.* **2019**, *19*, 227–247. [CrossRef]
36. Kaplan, M.; de la Torre, J. A blocked-CAT procedure for CD-CAT. *Appl. Psychol. Meas.* **2020**, *44*, 49–64. [CrossRef] [PubMed]
37. Wyse, A.E.; McBride, J.R. A framework for measuring the amount of adaptation of Rasch-based computerized adaptive tests. *J. Educ. Meas.* **2020**, *58*, 81–103. [CrossRef]
38. Wainer, H.; Kiely, G.L. Item clusters and computerized adaptive testing: A case for testlets. *J. Educ. Meas.* **1987**, *24*, 185–201. [CrossRef]
39. Wainer, H. Rescuing computerized testing by breaking Zipf's law. *J. Educ. Behav. Stat.* **2000**, *25*, 203–224. [CrossRef]
40. Chang, H.H. Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In *The SAGE Handbook of Quantitative Methodology for the Social Sciences*; Kaplan, D., Ed.; SAGE Publications, Inc.: Southend Oaks, CA, USA, 2004; pp. 118–135. [CrossRef]
41. Liu, Y.; Hau, K.T. Measuring motivation to take low-stakes large-scale test: New model based on analyses of "participant-own-defined" missingness. *Educ. Psychol. Meas.* **2020**, 1115–1144. [CrossRef] [PubMed]
42. Abdelfattah, F. The relationship between motivation and achievement in low-stakes examinations. *Soc. Behav. Personal. Int. J.* **2010**, *38*, 159–167. [CrossRef]
43. Wise, S.L.; DeMars, C.E. Examinee noneffort and the validity of program assessment results. *Educ. Assess.* **2010**, *15*, 27–41. [CrossRef]
44. Weiss, D.J. *Computerized Ability Testing, 1972–1975*; Final Report 150-343; Psychometric Methods Program, Department of Psychology, University of Minnesota: Minneapolis, MN, USA, 1976.
45. Martin, A.J.; Lazendic, G. Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *J. Educ. Psychol.* **2018**, *110*, 27–45. [CrossRef]
46. Arvey, R.D.; Strickland, W.; Drauden, G.; Martin, C. Motivational components of test taking. *Pers. Psychol.* **1990**, *43*, 695–716. [CrossRef]
47. Ling, G.; Attali, Y.; Finn, B.; Stone, E.A. Is a computerized adaptive test more motivating than a fixed-item test? *Appl. Psychol. Meas.* **2017**, *41*, 495–511. [CrossRef]
48. Asseburg, R.; Frey, A. Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychol. Test Assess. Model.* **2013**, *55*, 92–104.
49. Betz, N.E.; Weiss, D.J. *Psychological Effects of Immediate Knowledge of Results and Adaptive Ability Testing*; Research Report 76-4; Psychometric Methods Program, Department of Psychology, University of Minnesota: Minneapolis, MN, USA, 1976.
50. Wise, S.L. The utility of adaptive testing in addressing the problem of unmotivated examinees. *J. Comput. Adapt. Test.* **2014**, *2*, 1–17. [CrossRef]
51. Kimura, T. The impacts of computer adaptive testing from a variety of perspectives. *J. Educ. Eval. Health Prof.* **2017**, *14*, 1–5. [CrossRef]
52. Colwell, N.M. Test anxiety, computer-adaptive testing, and the common core. *J. Educ. Train. Stud.* **2013**, *1*, 50–60. [CrossRef]
53. Frey, A.; Hartig, J.; Moosbrugger, H. Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests [Effects of adaptive testing on test-taking motivation in the example of the Frankfurt adaptive test for measuring attention]. *Diagnostica* **2009**, *55*, 20–28. [CrossRef]
54. Pitkin, A.K.; Vispoel, W.P. Differences between self-adapted and computerized adaptive tests: A meta-analysis. *J. Educ. Meas.* **2001**, *38*, 235–247. [CrossRef]
55. Tonidandel, S.; Quiñones, M.A.; Adams, A.A. Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *J. Appl. Psychol.* **2002**, *87*, 320–332. [CrossRef]
56. Häusler, J.; Sommer, M. The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychol. Sci. Q.* **2008**, *50*, 75–87.
57. Ortner, T.M.; Caspers, J. Consequences of test anxiety on adaptive versus fixed item testing. *Eur. J. Psychol. Assess.* **2011**, *27*, 157–163. [CrossRef]
58. Lu, H.; Hu, Y.P.; Gao, J.J. The effects of computer self-efficacy, training satisfaction and test anxiety on attitude and performance in computerized adaptive testing. *Comput. Educ.* **2016**, *100*, 45–55. [CrossRef]
59. Hembree, R. Correlates, causes, effects, and treatment of test anxiety. *Rev. Educ. Res.* **1988**, *58*, 47–77. [CrossRef]
60. Wise, S.L. *Examinee Issues in CAT*; National Council on Measurement: Chicago, IL, USA, 1997.
61. O'Reilly, T.; Sabatini, J. Reading for understanding: How performance moderators and scenarios impact assessment design. *ETS Res. Bull. Ser.* **2013**, *2013*, i-47. [CrossRef]
62. Brown, S.M.; Walberg, H.J. Motivational effects on test scores of elementary students. *J. Educ. Res.* **1993**, *86*, 133–136. [CrossRef]

63. Wolf, L.F.; Smith, J.K. The consequence of consequence: Motivation, anxiety, and test performance. *Appl. Meas. Educ.* **1995**, *8*, 227–242. [CrossRef]
64. Mittelhaëuser, M.A.; Béuin, A.A.; Sijtsma, K. The effect of differential motivation on IRT linking. *J. Educ. Meas.* **2015**, *52*, 339–358. [CrossRef]
65. Finn, B. Measuring motivation in low-stakes assessments. *ETS Res. Bull. Ser.* **2015**, *2015*, 1–17. [CrossRef]
66. Stocking, M.L. Revising item responses in computerized adaptive tests: A comparison of three models. *Appl. Psychol. Meas.* **1997**, *21*, 129–142. [CrossRef]
67. Vispoel, W.P.; Rocklin, T.R.; Wang, T. Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Appl. Meas. Educ.* **1994**, *7*, 53–79. [CrossRef]
68. Papanastasiou, E.C.; Reckase, M.D. A "rearrangement procedure" for scoring adaptive tests with review options. *Int. J. Test.* **2007**, *7*, 387–407. [CrossRef]
69. Olea Díaz, J.; Revuelta Menéndez, J.; Ximénez, C.; Abad García, F.J. Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicológica* **2000**, *21*, 157–173.
70. Lunz, M.E.; Bergstrom, B.A.; Wright, B.D. The effect of review on student ability and test efficiency for computerized adaptive tests. *Appl. Psychol. Meas.* **1992**, *16*, 33–40. [CrossRef]
71. Lunz, M.E.; Bergstrom, B.A. An empirical study of computerized adaptive test administration conditions. *J. Educ. Meas.* **1994**, *31*, 251–263. [CrossRef]
72. Stone, E.; Davey, T. Computer-adaptive testing for students with disabilities: A review of the literature. *ETS Res. Bull. Ser.* **2011**, *2011*, i-24. [CrossRef]
73. Vispoel, W.P.; Rocklin, T.R.; Wang, T.; Bleiler, T. Can examinees use a review option to obtain positively biased ability estimates on a computerized adaptive test? *J. Educ. Meas.* **1999**, *36*, 141–157. [CrossRef]
74. Papanastasiou, E.C. Item review and the rearrangement procedure: Its process and its results. *Educ. Res. Eval.* **2005**, *11*, 303–321. [CrossRef]
75. Cui, Z.; Liu, C.; He, Y.; Chen, H. Evaluation of a new method for providing full review opportunities in computerized adaptive testing-computerized adaptive testing with salt. *J. Educ. Meas.* **2018**, *55*, 582–594. [CrossRef]
76. Wainer, H. Some practical considerations when converting a linearly administered test to an adaptive format. *Educ. Meas. Issues Pract.* **1993**, *12*, 15–20. [CrossRef]
77. Wise, S.L. *A Critical Analysis of the Arguments for and against Item Review in Computerized Adaptive Testing*; National Council on Measurement in Education: Chicago, IL, USA, April 1996.
78. Stone, G.E.; Lunz, M.E. The effect of review on the psychometric characteristics of computerized adaptive tests. *Appl. Meas. Educ.* **1994**, *7*, 211–222. [CrossRef]
79. Wang, S.; Fellouris, G.; Chang, H.H. Computerized adaptive testing that allows for response revision: Design and asymptotic theory. *Stat. Sin.* **2017**, 1987–2010. [CrossRef]
80. Wang, S.; Fellouris, G.; Chang, H.H. Statistical foundations for computerized adaptive testing with response revision. *Psychometrika* **2019**, *84*, 375–394. [CrossRef] [PubMed]
81. Lin, Z.; Chen, P.; Xin, T. The block item pocket method for reviewable multidimensional computerized adaptive testing. *Appl. Psychol. Meas.* **2020**, *45*, 22–36. [CrossRef] [PubMed]
82. Han, K.T. Item pocket method to allow response review and change in computerized adaptive testing. *Appl. Psychol. Meas.* **2013**, *37*, 259–275. [CrossRef]
83. Van der Linden, W.J.; Jeon, M.; Ferrara, S. A paradox in the study of the benefits of test-item review. *J. Educ. Meas.* **2011**, *48*, 380–398. [CrossRef]
84. Vispoel, W.P. Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *J. Educ. Meas.* **1998**, *35*, 328–345. [CrossRef]
85. Zwick, R.; Bridgeman, B. Evaluating validity, fairness, and differential item functioning in multistage testing. In *Computerized Multistage Testing: Theory and Applications*; Yan, A., von Davier, A.A., Lewis, C., Eds.; CRC Press: New York, NY, USA, 2014; pp. 271–300. [CrossRef]
86. Wise, S.L.; Kingsbury, G.G. Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica* **2000**, *21*, 135–155.
87. Kingsbury, G. *Item Review and Adaptive Testing*; National Council on Measurement in Education: New York, NY, USA, 1996.
88. Green, B.F.; Bock, R.D.; Humphreys, L.G.; Linn, R.L.; Reckase, M.D. Technical guidelines for assessing computerized adaptive tests. *J. Educ. Meas.* **1984**, *21*, 347–360. [CrossRef]
89. Svetina, D.; Liaw, Y.L.; Rutkowski, L.; Rutkowski, D. Routing strategies and optimizing design for multistage testing in international large-scale assessments. *J. Educ. Meas.* **2019**, *56*, 192–213. [CrossRef]
90. Kim, S.; Moses, T.; Yoo, H.H. Effectiveness of item response theory (IRT) proficiency estimation methods under adaptive multistage testing. *ETS Res. Bull. Ser.* **2015**, *2015*, 1–19. [CrossRef]
91. Yan, D.; von Davier, A.A.; Lewis, C. *Computerized Multistage Testing: Theory and Applications*; CRC Press: New York, NY, USA, 2014. [CrossRef]
92. Yamamoto, K.; Shin, H.J.; Khorramdel, L. Multistage adaptive testing design in international large-scale assessments. *Educ. Meas. Issues Pract.* **2018**, *37*, 16–27. [CrossRef]

93. Yamamoto, K.; Khorramdel, L. Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychol. Test Assess. Model.* **2018**, *60*, 347–368.
94. Berger, M.P. A general approach to algorithmic design of fixed-form tests, adaptive tests, and testlets. *Appl. Psychol. Meas.* **1994**, *18*, 141–153. [CrossRef]
95. Stark, S.; Chernyshenko, O.S. Multistage testing: Widely or narrowly applicable? *Appl. Meas. Educ.* **2006**, *19*, 257–260. [CrossRef]
96. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Pædagogiske Institut: Copenhagen, Denmark, 1960.
97. Holland, P.W. On the sampling theory roundations of item response theory models. *Psychometrika* **1990**, *55*, 577–601. [CrossRef]
98. San Martin, E.; De Boeck, P. What do you mean by a difficult item? On the interpretation of the difficulty parameter in a Rasch model. In *The 78th Annual Meeting of the Psychometric Society, Springer Proceedings in Mathematics & Statistics*; Quantitative Psychology Research; Millsap, R.E., Bolt, D.M., van der Ark, L.A., Wang, W.C., Eds.; Springer: New York, NY, USA, 2015; pp. 1–14. [CrossRef]
99. Molenaar, I.W. Some background for item response theory and the Rasch model. In *Rasch Models*; Fischer, G.H., Molenaar, I., Eds.; Springer: New York, NY, USA, 1995; pp. 3–14. [CrossRef]
100. De Boeck, P. Random item IRT models. *Psychometrika* **2008**, *73*, 533–559. [CrossRef]
101. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46*, 443–459. [CrossRef]
102. Bock, R.D.; Lieberman, M. Fitting a response model for n dichotomously scored items. *Psychometrika* **1970**, *35*, 179–197. [CrossRef]
103. Thissen, D. Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika* **1982**, *47*, 175–186. [CrossRef]
104. Andersen, E.B. The numerical solution of a set of conditional estimation equations. *J. R. Stat. Soc. Ser. B (Methodological)* **1972**, *34*, 42–54. [CrossRef]
105. Andersen, E.B. *Conditional Inference and Models for Measuring*; Mentalhygiejnisk Forlag: København, Denmark, 1973.
106. Wang, C.; Chen, P.; Jiang, S. Item calibration methods with multiple subscale multistage testing. *J. Educ. Meas.* **2019**, *57*, 3–28. [CrossRef]
107. Eggen, T.J.H.M.; Verhelst, N.D. Item calibration in incomplete testing designs. *Psicológica* **2011**, *32*, 107–132.
108. Glas, C.A.W. The Rasch model and multistage testing. *J. Educ. Stat.* **1988**, *13*, 45–52. [CrossRef]
109. Zwitser, R.J.; Maris, G. Conditional statistical inference with multistage testing designs. *Psychometrika* **2015**, *80*, 65–84. [CrossRef]
110. Xu, X.; von Davier, M. Fitting the structured general diagnostic model to NAEP data. *ETS Res. Bull. Ser.* **2008**, *2008*, i–18. [CrossRef]
111. Kubinger, K.D.; Steinfeld, J.; Reif, M.; Yanagida, T. Biased (conditional) parameter estimation of a Rasch model calibrated item pool administered according to a branched testing design. *Psychol. Test Assess. Model.* **2012**, *52*, 450–460.
112. Mislevy, R.J.; Sheehan, K.M. The role of collateral information about examinees in item parameter estimation. *Psychometrika* **1989**, *54*, 661–679. [CrossRef]
113. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
114. Holland, P.W.; Thayer, D.T. Univariate and bivariate loglinear models for discrete test score distributions. *J. Educ. Behav. Stat.* **2000**, *25*, 133–183. [CrossRef]
115. Casabianca, J.M.; Lewis, C. IRT item parameter recovery with marginal maximum likelihood estimation using smoothing models. *J. Educ. Behav. Stat.* **2015**, *40*, 547–578. [CrossRef]
116. Von Davier, M. A general diagnostic model applied to language testing data. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 287–307. [CrossRef]
117. Casabianca, J.M.; Junker, B.W. Estimating the latent trait distribution with loglinear smoothing models. In *New Developments in Quantitative Psychology*; Millsap, R.E., van der Ark, L.A., Bolt, D.M., Woods, C.M., Eds.; Springer: New York, NY, USA, 2013; pp. 415–425. [CrossRef]
118. Casabianca, J.M. Loglinear Smoothing for the Latent Trait Distribution: A Two-Tiered Evaluation. Ph.D. Thesis, Fordham University, Bronx, NY, USA, 2011.
119. Fischer, G.H. *Einführung in die Theorie Psychologischer Tests: Grundlagen und Anwendungen [Introduction into Theory of Psychological Tests]*; Huber: Berne, Switzerland, 1974.
120. Formann, A.K. A note on the computation of the second-order derivatives of the elementary symmetric functions in the Rasch model. *Psychometrika* **1986**, *51*, 335–339. [CrossRef]
121. Verhelst, N.D.; Glas, C.; van der Sluis, A. Estimation problems in the Rasch model: The basic symmetric functions. *Comput. Stat. Q.* **1984**, *1*, 245–262.
122. Liou, M. More on the computation of higher-order derivatives of the elementary symmetric functions in the Rasch model. *Appl. Psychol. Meas.* **1994**, *18*, 53–62. [CrossRef]
123. Eggen, T.J.H.M.; Verhelst, N.D. Loss of information in estimating item parameters in incomplete designs. *Psychometrika* **2006**, *71*, 303–322. [CrossRef]
124. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2020. Available online: https://www.R-project.org/ (accessed on 1 February 2020).
125. Bechger, T.; Koops, J.; Partchev, I.; Maris, G. *dexterMST: CML Calibration of Multi Stage Tests*; R Package Version 0.9.0; R Core Team: Vienna, Austria, 2020. Available online: https://CRAN.R-project.org/package=dexterMST (accessed on 20 September 2020).

126. Steinfeld, J.; Robitzsch, A. *tmt: Estimation of the Rasch Model for Multistage Tests*; R Package Version 0.2.1-0; R Core Team: Vienna, Austria, 2020. Available online: https://CRAN.R-project.org/package=tmt (accessed on 20 September 2020).
127. Casabianca, J.; Xu, X.; Jia, Y.; Lewis, C. *Estimation of Item Parameters When the Underlying Latent Trait Distribution of Test Takers Is Nonnormal*; National Council on Measurement in Education: Denver, CO, USA, April 2010.
128. Woods, C.M. Ramsay-curve item response theory for the three-parameter logistic item response model. *Appl. Psychol. Meas.* **2008**, *32*, 447–465. [CrossRef]
129. Woods, C.M.; Lin, N. Item response theory with estimation of the latent density using Davidian curves. *Appl. Psychol. Meas.* **2009**, *33*, 102–117. [CrossRef]
130. Woods, C.M.; Thissen, D. Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika* **2006**, *71*, 281–301. [CrossRef] [PubMed]
131. Smits, N.; Öğreden, O.; Garnier-Villarreal, M.; Terwee, C.B.; Chalmers, R.P. A study of alternative approaches to non-normal latent trait distributions in item response theory models used for health outcome measurement. *Stat. Methods Med Res.* **2020**, *29*, 1030–1048. [CrossRef] [PubMed]
132. Karadavut, T.; Cohen, A.S.; Kim, S.H. Estimation of mixture Rasch models from skewed latent ability distributions. *Meas. Interdiscip. Res. Perspect.* **2020**, *18*, 215–241. [CrossRef]
133. Sen, S. Spurious latent class problem in the mixed Rasch model: A comparison of three maximum likelihood estimation methods under different ability distributions. *Int. J. Test.* **2018**, *18*, 71–100. [CrossRef]
134. Wang, C.; Su, S.; Weiss, D.J. Robustness of parameter estimation to assumptions of normality in the multidimensional graded response model. *Multivar. Behav. Res.* **2018**, *53*, 403–418. [CrossRef]
135. Zwinderman, A.H.; van den Wollenberg, A.L. Robustness of marginal maximum likelihood estimation in the Rasch model. *Appl. Psychol. Meas.* **1990**, *14*, 73–81. [CrossRef]
136. Eysenck, H.; Eysenck, S. *Eysenck Personality Questionnaire–Revised*; Hodder and Stoughton: London, UK, 1991.
137. Wall, M.M.; Park, J.Y.; Moustaki, I. IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Appl. Psychol. Meas.* **2015**, *39*, 583–597. [CrossRef]
138. Micceri, T. The unicorn, the normal curve, and other improbable creatures. *Psychol. Bull.* **1989**, *105*, 156. [CrossRef]
139. Ho, A.D.; Yu, C.C. Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educ. Psychol. Meas.* **2015**, *75*, 365–388. [CrossRef]
140. Sass, D.; Schmitt, T.; Walker, C. Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Appl. Meas. Educ.* **2008**, *21*, 65–88. [CrossRef]
141. Mair, P.; Hatzinger, R.; Maier, M.J. *eRm: Extended Rasch Modeling*; R package version 1.0-2; R Core Team: Vienna, Austria, 2020. Available online: https://CRAN.R-project.org/package=eRm (accessed on 20 September 2020).
142. Zeileis, A.; Strobl, C.; Wickelmaier, F.; Komboz, B.; Kopf, J.; Schneider, L.; Debelak, R. *Psychotools: Infrastructure for Psychometric Modeling*; R package version 0.6-1; R Core Team: Vienna, Austria, 2020. Available online: https://CRAN.R-project.org/package=psychotools (accessed on 20 September 2020).
143. Robitzsch, A.; Steinfeld, J. *Immer: Item Response Models for Multiple Ratings, 2018*; R package version 1.1-35; R Core Team: Vienna, Austria, 2020. Available online: https://CRAN.R-project.org/package=immer (accessed on 20 September 2020).
144. Rizopoulos, D. ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *J. Stat. Softw.* **2006**, *17*, 1–25. [CrossRef]
145. Robitzsch, A. *sirt: Supplementary Item Response Theory Models*; R Package Version 3.9.4; R Core Team: Vienna, Austria, 2020. Available online: https://CRAN.R-project.org/package=sirt (accessed on 20 September 2020).
146. Chalmers, R.P. mirt: A Multidimensional Item Response Theory Package for the R Environment. *J. Stat. Softw.* **2012**, *48*, 1–29. [CrossRef]
147. Robitzsch, A.; Kiefer, T.; Wu, M. *TAM: Test Analysis Modules*; R Package Version 3.6-45; R Core Team: Vienna, Austria, 2020. Available online: https://CRAN.R-project.org/package=TAM (accessed on 20 September 2020).
148. Molenaar, I.W. Estimation of item parameters. In *Rasch Models*; Fischer, G.H., Molenaar, I., Eds.; Springer: New York, NY, USA, 1995; pp. 39–51. [CrossRef]
149. Robitzsch, A. A comparison of estimation methods for the Rasch model. In *Book of Short Papers—SIS 2021*; Perna, C., Salvati, N., Spagnolo, F. S., Eds.; Pearson: London, UK, 2021; pp. 157–162.
150. Rasch, G. On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements. In *The Danish Year-Book of Philosophy*; Blegvad, M., Ed.; Munksgaard: Copenhagen, Denmark, 1977; pp. 58–94.
151. Rasch, G. An informal report on a theory of objectivity in comparisons. Psychological measurement theory. In *Proceedings of the NUFFIC International Summer Session in Science at "Het Oude Hof" The Hague, Psychological Measurement Theory*; van der Kamp, L.J.T., Vlek, C.A.J., Eds.; University of Leiden: Leiden, The Netherlands, 1967.
152. Draxler, C. Bayesian conditional inference for Rasch models. *AStA Adv. Stat. Anal.* **2018**, *102*, 245–262. [CrossRef]