*Article*

Applied Psychological Measurement
2021, Vol. 45(6) 407–422

DOI: 10.1177/01466216211008530
journals.sagepub.com/home/apm
SAGE

# On the Speed Sensitivity Parameter in the Lognormal Model for Response Times and Implications for High-Stakes Measurement Practice

**Benjamin Becker[1], Dries Debeer[2], Sebastian Weirich[1], and Frank Goldhammer[3,4]**

## Abstract

In high-stakes testing, often multiple test forms are used and a common time limit is enforced. Test fairness requires that ability estimates must not depend on the administration of a specific test form. Such a requirement may be violated if speededness differs between test forms. The impact of not taking speed sensitivity into account on the comparability of test forms regarding speededness and ability estimation was investigated. The lognormal measurement model for response times by van der Linden was compared with its extension by Klein Entink, van der Linden, and Fox, which includes a speed sensitivity parameter. An empirical data example was used to show that the extended model can fit the data better than the model without speed sensitivity parameters. A simulation was conducted, which showed that test forms with different average speed sensitivity yielded substantial different ability estimates for slow test takers, especially for test takers with high ability. Therefore, the use of the extended lognormal model for response times is recommended for the calibration of item pools in high-stakes testing situations. Limitations to the proposed approach and further research questions are discussed.

## Keywords

test assembly, speededness, item response theory, high-stakes assessment

In high-stakes assessments such as college administration tests (e.g., SAT; College Board, 2016) or language proficiency tests (e.g., TOEFL; Educational Testing Service [ETS], 2020), important consequences result from test scores, such as admission to university or other

[1]Humboldt University of Berlin, Germany
[2]KU Leuven, Belgium
[3]DIPF – Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany
[4]Centre for International Student Assessment Germany (ZIB)

**Corresponding Author:**
Benjamin Becker, Institute for Educational Quality Improvement, Humboldt University of Berlin, Unter den Linden 6, 10099 Berlin, Germany.
Email: b.becker@iqb.hu-berlin.de

educational programs. The high-stakes connected to the test outcome have important implications for the design and analysis of the respective tests. First, to increase test security, often multiple parallel test forms are used. This prevents cheating during testing sessions with multiple test takers and sharing knowledge about the test by former test takers (Luecht & Sireci, 2011). Second, for reasons of fairness, testing conditions are standardized across test takers and test occasions. For instance, the time limit for the test is equal regardless of the test form. Third, due to the high-stakes, test takers are often assumed to be highly motivated. Therefore, missing responses are commonly considered informative, that is, they are scored as incorrect responses. This scoring rule is communicated to test takers, to prevent test takers from strategically not responding to items they feel unable to provide a correct response to. Ignoring missing values as a scoring rule could incentivize test takers to omit these items and thereby lead to biased and unfair ability estimates.

When multiple tests forms are used, they are often required to be parallel, which in the strict sense means that for every test taker, the test forms have the same true score and the same error variance (Lord & Novick, 1986). Within an item response theory (IRT) framework where maximum likelihood is used to estimate ability, the expected ability estimate $E(\hat{\theta})$ and the expected standard error $E(\widehat{SE_\theta})$ for all test takers should be independent of the administered test form $z$, which corresponds to so-called weak parallelism (Samejima, 1977). When missing responses are scored as incorrect, differences in the speededness of the test forms can violate this requirement.[1] Imagine one test taker, working at a specific speed, and a test with two test forms A and B that only differ in their expected testing time for the specific test taker. The time limit for the test administration is 60 min and the expected total response time of the test taker on test form A is 60 min but 70 min on Test Form B. When confronted with test form B, the test taker has to choose from three strategies: (a) Work with the identical speed as on test form A and not reach the end of the test, (b) work with the identical speed as on test form A and omit items, or (c) work with increased speed and respond to all items in time. Missing responses resulting from (a) and (b) are scored as incorrect. Working with increased speed (c) usually leads to decreased accuracy (cf. the within-person speed–accuracy trade-off; Goldhammer, 2015). Hence, all strategies will result in a lower expected ability estimate on test form B compared with test form A. Combinations of the three strategies are also plausible but will have similar consequences on the ability estimate.

The example illustrates that the speededness of a test is an interaction of the time intensity of its items, the time limit set on the test, and the exerted working speed of the test taker (van der Linden, 2011b). As the speed level usually varies between persons, the degree of speededness of a test can also be expected to vary between persons. A fast and proficient test taker will score higher on a test with a time limit than an equally proficient but slower test taker who has to engage in one of the above-described strategies to deal with the insufficient time available. Consequently, however, the measured latent construct is no longer a pure ability measure, but a composite measure of speed and ability. Whether this is seen as a conceptual property of the test or a byproduct of the testing conditions differs. In this article, there are no assumptions made on the nature of speed differences between persons and to which degree they should affect ability measurement in high-stakes testing.[2] Instead, the article focuses on how to hold the level of speededness constant across all test forms within each individual test taker.

In the following section, the typical test assembly process and analysis that is commonly performed to obtain individual ability estimates in high-stakes assessments is briefly outlined. Based on this, the state-of-the-art approach to prevent differentially speeded test forms, which uses latent response time modeling, is described. An important shortcoming of this model is explained and a common model extension that mitigates this shortcoming is discussed.

## Assessment Framework

### Test Assembly

The common process of creating multiple parallel test forms contains the following steps (College Board, 2016; van der Linden, 2005): (a) developing items, (b) using items on a piloting sample (Piloting), (c) item parameter estimation (Calibration), and (d) assembly of items from an item pool to parallel test forms (Test Assembly). Criteria for the assembly of tests, besides test speededness, include the test information function, comparability of content, and similar distribution of item types (van der Linden, 2005). Due to the emergence of computer-administered testing, balancing speededness has become substantially easier. In this article, it is assumed that response times are available from a computer-administered piloting study.

### Ability Estimation

For the estimation of latent abilities, an often-used choice is the two-parameter logistic (2PL) model. As already described, it is assumed that missing responses are scored as incorrect. Throughout this article, the notation of Fox (2010) is adopted, denoting items as $k = 1, ..., j$ and persons as $i = 1, \ldots, n$, with correct responses denoted as $y_{ik} = 1$. In the 2PL model, the probability to solve an item $k$ correctly can be denoted as follows:

$$P(y_{ik} = 1 | \theta_i, a_k, b_k) = \frac{\exp(a_k\theta_i - b_k)}{1 + \exp(a_k\theta_i - b_k)}. \quad (1)$$

## Balancing Speededness

Several strategies have been proposed to balance speededness across the test forms of a test administration, for example, using observed response times from a piloting study (e.g., van der Linden, 2005). In the following section, the current state-of-the-art approach, which uses a latent measurement model for response times, is discussed.

### Lognormal Measurement Model

Recently, van der Linden (2011b) proposed the use of a lognormal latent measurement model for response times (van der Linden, 2006) for balancing speededness across test forms. The model assumes response times to be lognormally distributed and parameterizes these lognormal response times, $\ln \mathrm{RT}_{ik}$, as follows:

$$\ln \mathrm{RT}_{ik} = \lambda_k - \zeta_i + \epsilon_{ik}, \quad \text{with } \epsilon_{ik} \sim N\left(0, \sigma^2_{\epsilon_k}\right), \quad (2)$$

where $\lambda_k$ represents the *time intensity* of a specific item, whereas $\zeta_i$ represents the average speed with which a person works (*person speed parameter*). In addition, an item-specific residual variance $\sigma^2_{\epsilon_k}$ is estimated. As the model is parameterized with two item-specific parameters, this article refers to it as the *two-parameter lognormal model* (2PLN). In his study, van der Linden (2011a) proposed controlling the expected testing time, conditionally on the speed parameter, according to the 2PLN. He showed that this approach performed better than using observed response times to balance speededness across test forms (van der Linden, 2011b), as it, for example, also controls for differing variances in response times between items.

*Speed Discrimination.* According to the 2PLN, items can have different intercept parameters $\lambda_k$ and different residual variances $\sigma^2_{\epsilon_k}$. Furthermore, van der Linden (2006) introduced the inverse of the residual variance as the discrimination parameter $\alpha_k$:

$$\alpha_k = \frac{1}{\sigma^2_{\epsilon_k}}. \tag{3}$$

$\alpha_k$ thereby represents the precision of the response time distribution (Molenaar et al., 2015). This article, however, to avoid confusion, only refers to the residual variance, and not to its inverse. Compared with models from confirmatory factor analysis, the 2PLN resembles a tau-equivalent measurement model (Brown, 2006) for log response times. This means the model lacks a slope parameter and therefore, speaking in terms of more generalized models, assumes that the slope parameter is equal across all items or indicators. This equals the assumption that items with equal residual variances correlate all equally strong with the measured latent construct. Conceptually speaking for response time modeling, this means that the 2PLN assumes that items do not differ in their sensitivity to speed differences across persons.

This article, however, will argue that items can differ in the extent to which they are sensitive to speed differences, and that this variability across items needs to be taken into account when assembling test forms that should have equal speededness for each test taker. In the next section, an extension of the lognormal measurement model for response times which allows differences in speed sensitivity across items is discussed.

## Extension of the Lognormal Measurement Model

Klein Entink, Fox, and van der Linden (2009) proposed an extension of the 2PLN that this article refers to as the *three-parameter lognormal model* (3PLN). It introduces a slope parameter $\phi_k$. This measurement model resembles a congeneric measurement model for log-transformed response times in confirmatory factor analysis (Brown, 2006):

$$\ln \mathrm{RT}_{ik} = \lambda_k - \phi_k \zeta_i + \epsilon_{ik}, \quad \text{with} \quad \epsilon_{ik} \sim N\left(0, \sigma^2_{\epsilon_k}\right). \tag{4}$$

Conceptually, the parameter $\phi_k$ allows for individual items being more sensitive to speed differences between test takers than other items. To avoid confusion with the $\alpha_k$ parameter that van der Linden (2006) labels as a discrimination parameter in the 2PLN, the authors will use the term *speed sensitivity* to refer to $\phi_k$ throughout this article.

*Difference between the 2PLN and the 3PLN.* There has been some confusion around the 2PLN and the 3PLN and the meaning of their respective item parameters in the literature.[3] It is important to note that the 2PLN and the 3PLN models are not equivalent formulations of the same model. This can be illustrated by comparing the model implicit correlations between the response times of two items $k$ and $l$ of the 2PLN and the 3PLN. For the 2PLN, this correlation is defined as follows:

$$\rho_{\mathrm{RT}_k, \mathrm{RT}_l} = \frac{\left[\exp\left(\sigma^2_\zeta\right) - 1\right]}{\sqrt{\left(\exp\left(\sigma^2_{\epsilon_k} + \sigma^2_\zeta\right) - 1\right)\left(\exp\left(\sigma^2_{\epsilon_l} + \sigma^2_\zeta\right) - 1\right)}}. \tag{5}$$

In contrast, for the 3PLN, this correlation is defined as follows:

$$\rho_{\mathrm{RT}_k,\mathrm{RT}_l} = \frac{\left[\exp\left(\phi_k\phi_l\sigma_\zeta^2\right) - 1\right]}{\sqrt{\left(\exp\left(\sigma_{\epsilon_k}^2 + \phi_k^2\sigma_\zeta^2\right) - 1\right)\left(\exp\left(\sigma_{\epsilon_l}^2 + \phi_l^2\sigma_\zeta^2\right) - 1\right)}}. \quad (6)$$

For the derivation of both formulas, see Online Appendix A. For a similar remark on the model implicit covariances of the response times of two items, see Fox and Marianti (2016).

To illustrate the difference between the residual variance and the speed sensitivity parameter, Figure 1 shows response time distributions conditional on two different speed levels ($\zeta_1 = 1$, $\zeta_2 = -1$) for four different items. The left side of the figure shows the distributions for items with a high residual variance $\sigma_{\epsilon_k}^2 = 1.00$, and the right side for items with a low residual variance, $\sigma_{\epsilon_k}^2 = 0.33$. Furthermore, the upper half of the figure depicts the distributions for items with low speed sensitivity, $\phi_k = 0.3$, and the lower half items with high speed sensitivity, $\phi_k = 1$. The graphs illustrate how the residual variance controls the broadness of the distributions (and is strongly connected to the concept of reliability), whereas $\phi_k$ controls how far the medians of the response time distributions differ between persons with differing speed levels. An identical figure for the log-transformed response times can be seen in Online Appendix B.

As an illustration of the conceptual meaning of the speed sensitivity of items, consider the following two hypothetical math items with equal time intensity (e.g., $\lambda_1 = \lambda_2 = 4$). The first item embeds a simple task in a long text; the second item has no text to read, but requires a lengthy calculation. It seems plausible to assume that the second item is more sensitive to working speed specific to math items (e.g., $\phi_1 = 0.7$), because the calculation is longer. In contrast, the first item could be less sensitive to mathematical working speed because the response time mostly depends on the reading speed ($\phi_2 = 0.3$). As reading and mathematical literacy are assumed to be distinct constructs, this is plausibly also the case for reading and mathematical speed. The consequences for the Response Time Characteristic Curve, as also described in Fox (2010), can be seen in Online Appendix C. These two items would not lead to differences in response times for medium speed levels ($\zeta_k = 0$) but to substantial differences for slow ($\zeta_k = -1$) and fast test takers ($\zeta_k = 1$), with differences increasing with increasing deviation from $\zeta_k = 0$. For a test taker with $\zeta_i = -1$, the expected response times of the two example items are 73.70 and 109.95 s. As time pressure usually only occurs for slow participants, generally only differences in response times for slow but not for fast participants will be relevant for the estimation of ability in educational assessments.

## Hierarchical Framework

For model estimation in the context of test assembly, van der Linden (2011a) proposed embedding the lognormal latent measurement model for response times in a hierarchical framework (van der Linden, 2007). The resulting model assumes two latent dimensions, ability and speed, with common item and person parameter distributions. Conditional on these joint distributions, the model assumes independently distributed responses and response times. The framework benefits the estimation of the two dimensions, especially if the two dimensions are correlated (van der Linden et al., 2010). The joint person parameter distribution with either the 2PLN or the 3PLN is a multivariate normal distribution with

$$(\theta_i, \zeta_i) \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{P}, \boldsymbol{\Sigma}_\mathbf{P}). \quad (7)$$

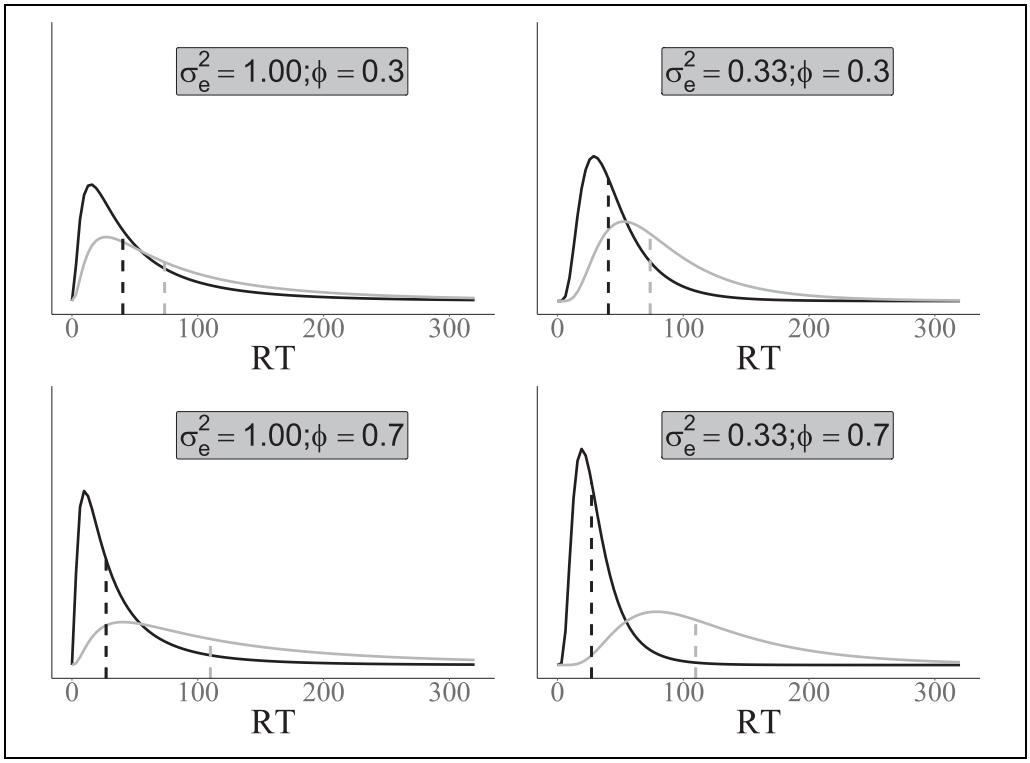


**Figure 1.** Conditional response time distributions for a fast speed level with $\zeta_1 = 1$ (black line) and a slow speed level with $\zeta_2 = -1$ (gray line) on four different items, all with $\lambda_k = 4$.
*Note.* Dotted lines indicate the medians of the corresponding distributions.

The joint item parameter distribution with the 2PLN together with a 2PL model for ability is also a multivariate normal distribution[4] with

$$(a_k, b_k, \lambda_k) \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{I}, \boldsymbol{\Sigma}_\mathbf{I}). \tag{8}$$

The joint item parameter distribution with the 3PLN together with a 2PL model for ability also includes $\phi_k$:

$$(a_k, b_k, \lambda_k, \phi_k) \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{I}, \boldsymbol{\Sigma}_\mathbf{I}). \tag{9}$$

## Research Questions

The questions arise, whether (a) the hierarchical framework with the 3PLN as a measurement model for response times fits empirical response time data better than the hierarchical framework with the 2PLN and, if this is the case, (b) what the consequences would be for ability estimation in high-stakes assessments. To the authors' knowledge, models with the 2PLN and the 3PLN have not yet been compared using data from educational competence tests. Moreover, there have only been a few comparisons using empirical data at all, so far focusing on intelligence tests (Goldhammer & Klein Entink, 2011), complex problem-solving tasks (Scherer et al., 2015), and mental rotation tasks (Debelak et al., 2014). In all three studies, the model with the 3PLN showed better fit than the model with the 2PLN according to the DIC (Spiegelhalter

et al., 2002). In addition, the hierarchical framework with the 3PLN has been applied to noneducational vocational credentialing high-stakes data (Fox & Marianti, 2017) and low-stakes data of chess tasks (Fox & Marianti, 2016). In both cases, substantial variance in the speed sensitivity parameter was found across the items. The aforementioned studies provide general evidence for the relevance of the proposed model extension. However, they do not focus on educational assessment data. Therefore, an empirical data analysis was conducted, in which the models with the 2PLN and the 3PLN were applied and compared with data from an educational assessment, to investigate whether items differ in their speed sensitivity. This analysis is discussed in the ''Empirical Data Analysis'' section.

If the appropriateness of the model extension indeed holds in educational competence testing and items vary in their speed sensitivity, those differences may also accumulate over test forms of educational high-stakes assessments. This could result in test forms that, despite having equal time intensities and similar average observed response times, differ in their sensitivity to speed differences and therefore in their conditional distributions of expected testing times. Especially the substantial differences in expected response times for slow test takers would be important, as they could lead to differences in ability estimates across test forms. In the section ''Simulation Study,'' the possible consequences of unbalanced test forms on ability estimation are investigated and described using simulated data from test forms with item properties as found in the empirical example.

## Empirical Data Analysis

### *Data Description*

For the empirical data analysis, data from the 2015 Programme of International Student Assessment (PISA, Organisation for Economic Co-operation and Development [OECD], 2016) were used, for which responses and response times on item level are publicly available. The competences measured by PISA resemble competences that are often assessed in high-stakes educational assessments. Note that it is not uncommon to calibrate items for a high-stakes context based on data from low-stakes conditions, when piloting in high-stakes conditions is cumbersome or impossible (e.g., College Board, 2016; ETS, 2020). In those situations, it is implicitly assumed that items function similarly in low- and high-stakes conditions. In that sense, the results of this empirical low-stakes data analysis also have implications for high-stakes assessments. The Canadian subsample was chosen because it is the largest among the 72 countries participating in PISA.

To avoid substantial numbers of missing responses by design, test booklets were analyzed separately and only the test takers who had worked on the respective booklet were included. In PISA 2015, every test form consisted of four booklets, and booklets were assembled to a whole of 66 different test forms in the computer-administered version. Returning to items within a booklet was only possible within the items sharing a common stimulus and otherwise prohibited. Response times were accumulated across multiple visits of the same item (OECD, 2016). All math booklets used in the assessment were analyzed (named ''M01''–''M05'' and ''M06ab''), which appeared each in overall eight different test forms, at every position twice. For simplicity, all polytomous items were dichotomized, scoring fully correct responses as correct and partially incorrect responses as incorrect. This resulted in data sets of 10 to 12 dichotomous items and 1,863 to 1,929 persons.

## Method

The software JAGS (Plummer, 2003) together with the R package rjags (Plummer, 2016) was used for model estimation. The hierarchical framework with both the 2PLN and the 3PLN was used to analyze the data set. In the actual analysis of the PISA data set, omitted responses are scored incorrect and number of not-reached responses is used as a manifest variable in the background model for the plausible value generation (OECD, 2016). Because the aim of this empirical example is the unbiased estimation of item parameters (as in an actual pilot study for a high-stakes assessment), all missing responses were treated as if the items were not administered to the corresponding persons, which is the recommended practice for estimating item parameters (Finch, 2008).

*Model estimation.* Priors were uninformative and chosen in correspondence to Fox (2010) and Pohl et al. (2019). An inverse Wishart distribution was used as a hyperprior for the distribution of the three ($b_k$, $a_k$, $\lambda_k$) or respectively four item parameters ($b_k$, $a_k$, $\lambda_k$, $\phi_k$). Further information on the prior distributions can be seen in Online Appendix D. The DIC was calculated and compared between the two models to assess model fit (Spiegelhalter et al., 2002). The posterior distributions of the speed sensitivity parameters and their mean and standard deviation were investigated.

## Results

Inspections of the Markov chain Monte Carlo (MCMC) chains were conducted using the R packages coda (Plummer et al., 2006) and rjags. Trace plots indicate good convergence for all parameters in both models in all data sets. The point estimates of the univariate potential scale reduction factors (Gelman & Rubin, 1992) for all parameters in all booklets were below 1.03 (95% upper confidence interval limits at or below 1.10) and below 1.05 (95% upper confidence interval limits at or below 1.19), for the models with, respectively, the 2PLN and the 3PLN. This indicates satisfactory convergence (Gelman & Shirley, 2011). The correlation of the person ability and person speed parameter ranged between $r_{\theta_i\zeta_i} = -.62$ in booklet ''M01'' and $r_{\theta_i\zeta_i} = -.49$ in booklet ''M02,'' indicating a medium negative relationship between ability and speed. Similar results have been reported and are often explained by the fact that test takers need more time if they actually solve an item (Debelak et al., 2014; Goldhammer & Klein Entink, 2011; Scherer et al., 2015). If test takers are not able to solve an item, they may guess and move on to the next item.

Regarding model fit, DIC indicated better fit with the 3PLN as a measurement model for all booklets (Online Appendix E). Table 2 shows the statistics for the resulting speed sensitivities for all booklets. The mean of speed sensitivities $M(\phi_k)$ within booklets ranged from to 0.37 to 0.47, whereas $SD(\phi_k)$ ranged from 0.32 to 0.36. The 95% highest posterior density (HPD) interval for the standard deviation $SD(\phi_k)$ excluded 0 for all booklets. These findings provide evidence that there was substantial variation in the speed sensitivity across items in the empirical data.

The correlations of the speed sensitivities with other item parameters were also investigated. Table 1 displays the means of the posterior distributions of these correlations. Speed sensitivity correlated low but consistently over all booklets with the time intensity parameter $\lambda_k$ and difficulty parameter $b_k$. There was more variation across booklets in the correlation with the discrimination parameter $a_k$, but correlations were still small or close to 0. The small correlations imply that the speed sensitivity parameter is largely independent from the other item parameters and would not be indirectly balanced if the other item parameters were balanced between test forms.

**Table 1.** Descriptive Statistics of Item Speed Sensitivity Within All Math Booklets.

| Booklet | $M(\phi)$ | $SD(\phi)$ | 95% HPD | Min($\phi$) | Max($\phi$) | $r_{\phi, b}$ | $r_{\phi, a}$ | $r_{\phi, \lambda}$ |
|---|---|---|---|---|---|---|---|---|
| M01 | 0.40 | 0.34 | [0.20, 0.47] | 0.15 | 0.75 | 0.28 | –0.09 | 0.21 |
| M02 | 0.37 | 0.36 | [0.21, 0.52] | 0.14 | 0.57 | 0.19 | 0.12 | 0.16 |
| M03 | 0.39 | 0.33 | [0.20, 0.46] | 0.29 | 0.53 | 0.13 | 0.02 | 0.08 |
| M04 | 0.42 | 0.34 | [0.20, 0.46] | 0.18 | 0.67 | 0.27 | 0.28 | 0.16 |
| M05 | 0.44 | 0.32 | [0.19, 0.44] | 0.25 | 0.66 | 0.21 | –0.01 | 0.18 |
| M06ab | 0.47 | 0.35 | [0.21, 0.48] | 0.19 | 0.69 | 0.30 | 0.26 | 0.19 |

*Note.* Descriptive statistics for speed sensitivity, including its mean $M(\phi_k)$, standard deviation $SD(\phi_k)$, the HPD interval for the standard deviation $SD(\phi_k)$, minimum (Min($\phi_k$)) and Maximum (Max($\phi_k$)), and correlations of speed sensitivity with the other item parameters. HPD = highest posterior density.

## Simulation Study

### Design

The performed empirical data analyses illustrate that it is plausible to assume differences between items regarding their speed sensitivity. Therefore, the question arises, how the fairness of test forms is affected if this speed sensitivity is not controlled for between test forms. Based on the findings and parameter distributions in the empirical analyses, a simulation study was conducted to investigate how differences in speed sensitivity across test forms affect ability estimates. The simulation study reflects the main stage of a high-stakes assessments in which item properties are known from prior piloting and the sole interest lies in person parameter estimation. Three test forms were created, each with 30 items. The item parameters for the first test form were drawn from a multivariate normal distribution. Means, variances, and covariances of the item parameters were set to be in accordance with the results obtained from the empirical data analysis (see Online Appendix F). $\phi_k$ and $a_k$ were truncated at 0. If an item parameter draw included any $\phi_k$ and $a_k$ smaller than 0, all item parameters were drawn again for this replication. The residual variance of the log response times was drawn from a univariate normal distribution with $\sigma^2_{\epsilon_k} \sim \mathcal{N}(\mu = 0.2, \sigma^2 = 0.1)$, also truncated at 0. This first test form, with $\mu(\phi_k) = 0.3$, is referred to as the *low speed sensitivity test form*. A second and third test form were created with identical item parameters but shifts in their average speed sensitivity, resulting in a *medium speed sensitivity test form* with $\mu(\phi_k) = 0.4$ and a *high speed sensitivity test form* with $\mu(\phi_k) = 0.7$. The difference in speed sensitivity between the first and second test form reflects a common difference between booklets, which can also be found in the empirical example. Therefore, the comparison between these booklets can be used to determine expected bias even if only a few test forms are assembled. The difference between the first and third test form reflects a more extreme but not implausible case.[5] This condition was chosen to illustrate the theoretically possible impact of differing speed sensitivities and potential bias if a large number of test forms is assembled. Person parameters were chosen to enable conclusions about the effect of the two differing test forms on all possible combinations of speed and ability. Therefore, 500 ability parameters from $\theta_i \sim \mathcal{N}(0, 1)$ were sampled and combined with four different levels of speed, $\zeta_i = [-1; -0.5; 0.5; 1]$. This resulted in a complete sample of $n = 2,000$ test takers across the four speed subgroups. Responses and response times of the complete sample working on both test forms were simulated according to the hierarchical framework with the 3PLN and the 2PL. The time limit was set to 65 min (3,900 s) to introduce a reasonable amount of not-reached items into the simulation. Overall, 500 replications were conducted.

## Method

Person abilities were estimated according to the 2PL, with known item parameters using the weighted likelihood estimator (WLE; Warm, 1989) via the R package TAM (Robitzsch et al., 2017). Not-reached items were scored as incorrect. This approach reflects a high-stakes assessment, in which item parameters are obtained from a previously conducted calibration study and ability estimation is the focus (without specifically considering speed in the estimation). Numbers of not-reached items and estimated ability were compared for the four different speed groups between the three test forms.

## Results

As can be predicted from the response time measurement model in Equation 4 and the response time characteristic curves described in the introduction, differences in cumulative response times between the three test forms were most severe for the fastest and slowest participants (Table 2).[6] The fastest subgroup was much faster than the time limit of 3,900 s, with means of 1,310.08 and 1,953.53 s for the high and the low speed sensitivity test forms. In contrast, the slowest subgroup working on the high speed sensitivity test form was, on average, substantially slower than the time limit, with a mean of 5,419.48 s. In the faster subgroups, the differences in testing time did not result in different numbers of not-reached items because for all test forms the testing times were well below the time limit. For the slowest participants, however, the medium and high speed sensitivity test form led to substantially more not-reached items than the low speed sensitivity test form. Detailed numbers for items not-reached on average can be seen in Table 2 and are depicted in Online Appendix G for a single replication.

These differences in number of not-reached items also resulted in differences in ability estimates, mainly for the slowest subgroup. For them, the average difference in ability estimation between the test forms with low and medium speed sensitivity was 0.09 and 0.51 between the test forms with low and high speed sensitivity. Higher average speed sensitivity resulted in substantially lower ability estimates. A difference of 0.51 in the ability logit for a test taker with a true ability $\theta_i = 0$ is equal to a drop from the 50th ability percentile to the 32th ability percentile. However, differences in ability estimation were not homogeneous within the slowest subgroup. Especially slow participants with high ability had substantially different ability estimates depending on the test form (see the upper left graph in Figure 2). Average differences in ability estimation were also calculated for the quantile including only the 25% most able test takers, resulting in differences in ability estimation of 0.15 between the low and medium and 0.78 between the low and high speed sensitive test forms. This was to be expected because for slow but high ability test takers there are many not-reached items (scored as incorrect) that they could have answered correctly under sufficient time conditions. This is not the case for slow and low-ability test takers, for which only minor differences in estimated abilities across the test forms occurred. Furthermore, differences in speed sensitivities between test forms resulted in higher root mean square errors (RMSEs) and lower correlations between estimated and true ability parameters (see Table 2) for more speed sensitive test forms.

To conclude, the simulation shows that differences in speed sensitivity between test forms can lead to substantial differences in ability estimates especially for slow and able test takers. This finding is independent from whether speed is seen as a nuisance parameter or part of the construct to be measured. Furthermore, if speed is seen as a nuisance parameter, the high speed sensitivity test forms lead to a more biased and less precise ability measurement. If speed is seen as a substantial part of the construct to be measured, differences between true and

**Table 2.** Test Statistics per Test Form and per Speed Group, Averaged Across All Replications.

| Test form | $\zeta_i$ | *M*(RT) | *SD*(RT) | *M*(mis) | *SD*(mis) | cor($\hat{\theta}, \theta$) | RMSE | $M(\Delta_{\theta})$ |
|---|---|---|---|---|---|---|---|---|
| Low $\phi$ | Slowest | 3,636.44 | 371.73 | 0.02 | 0.04 | 0.90 | 0.47 | –0.04 |
| Low $\phi$ | Slow | 3,102.47 | 313.97 | 0.00 | 0.01 | 0.91 | 0.45 | –0.00 |
| Low $\phi$ | Fast | 2,275.08 | 228.36 | 0.00 | 0.00 | 0.91 | 0.45 | –0.00 |
| Low $\phi$ | Fastest | 1,953.53 | 195.04 | 0.00 | 0.00 | 0.91 | 0.45 | –0.00 |
| Medium $\phi$ | Slowest | 4,019.63 | 410.14 | 0.06 | 0.07 | 0.89 | 0.51 | –0.12 |
| Medium $\phi$ | Slow | 3,261.12 | 329.52 | 0.00 | 0.02 | 0.91 | 0.46 | –0.01 |
| Medium $\phi$ | Fast | 2,163.99 | 217.06 | 0.00 | 0.00 | 0.91 | 0.45 | –0.00 |
| Medium $\phi$ | Fastest | 1,767.68 | 176.80 | 0.00 | 0.00 | 0.91 | 0.45 | –0.00 |
| High $\phi$ | Slowest | 5,424.72 | 553.29 | 0.29 | 0.08 | 0.82 | 0.83 | –0.55 |
| High $\phi$ | Slow | 3,788.69 | 382.56 | 0.03 | 0.06 | 0.90 | 0.48 | –0.06 |
| High $\phi$ | Fast | 1,862.25 | 186.12 | 0.00 | 0.00 | 0.91 | 0.45 | –0.00 |
| High $\phi$ | Fastest | 1,310.08 | 130.71 | 0.00 | 0.00 | 0.91 | 0.45 | –0.00 |

*Note.* Descriptive statistics are depicted for mean cumulative response times *M*(RT) and the corresponding standard deviation *SD*(RT), mean proportion of missings *M*(mis), the corresponding standard deviation *SD*(mis), correlation between true and estimated ability cor($\hat{\theta}, \theta$), root mean square error (RMSE), and average difference between true and estimated ability $M(\Delta_{\theta})$.

estimated ability are in fact desirable for slow test takers, however should be identical across test forms.

## Discussion

High-stakes assessments often require multiple test forms with equal speededness at the level of the test taker. So far, the use of average response times and the use of the lognormal measurement model for response times by van der Linden (2006) have been proposed as strategies to control speededness across test forms (van der Linden, 2011b). In this article, the 2PLN model was compared to the model extension of the 3PLN by Klein, Fox, and van der Linden (2009), which introduces a speed sensitivity parameter into the measurement model. It was investigated which measurement model, embedded in the hierarchical framework by van der Linden (2007), fits empirical competence data better. Indeed, the 3PLN showed better model fit and the estimated speed sensitivity parameters varied substantially across items. This implies that balancing test forms using either observed response times or the item parameters from the 2PLN can lead to unbalanced speed sensitivity across test forms. Moreover, the simulation study shows that when missing responses are treated as incorrect (a standard practice in high-stakes assessments), differences in speed sensitivity between test forms can lead to severe differences in ability estimation. Especially slow test takers with a high ability were affected because they had increased numbers of not-reached items in the test forms that had higher speed sensitivities.

The issue of differential speed sensitivity can also be illustrated from an alternative perspective: As stated before, it is assumed that high-stakes tests usually are speeded power tests and therefore that the ability measured in the test is a composite measure of ability and speed. However, this composition changes between test forms if the test forms differ in their speed sensitivity. If a test form has a high speed sensitivity and a time limit induces time pressure for a certain speed level, the proportion of speed in the composite measure can be considered quite high. If in the same scenario a test form has low speed sensitivity, however, the proportion of speed in the composite measure for this test form will be rather low. This study argues that the

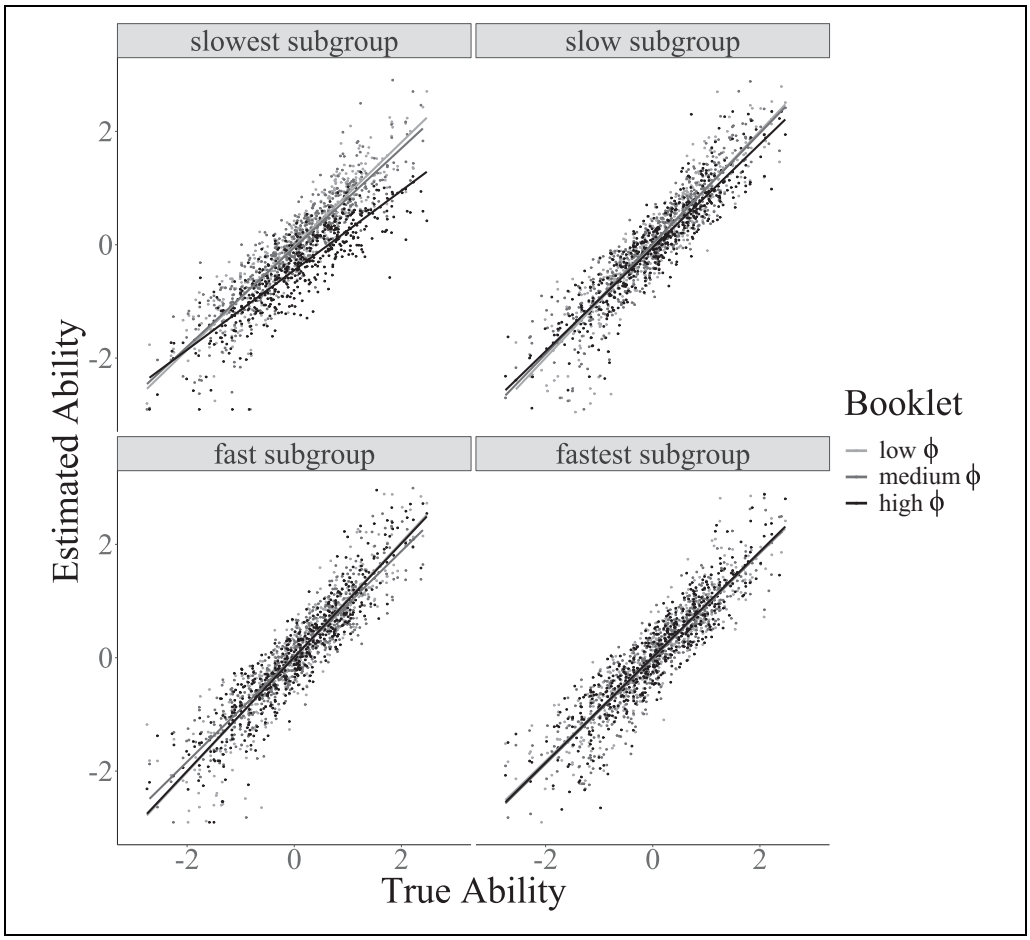


**Figure 2.** True and estimated ability for the low and high speed sensitivity test form, across the four subgroups.

*Note.* Results shown for a randomly selected single replication.

influence of speed on the ability estimation has to be the same across test forms within each speed level.

## *Practical Implications*

The following conclusions are drawn regarding the practice of assembling test forms for educational high-stakes assessments: Right now, the use of the hierarchical framework with the 2PLN is the state-of-the-art approach when balancing test forms. However, the findings of this study suggest that only when (a) the model with the 2PLN proves to better fit the data than the model with the 3PLN (e.g., using DIC) or (b) the 3PLN shows low variation in the speed sensitivity parameter across items, this approach should be considered sufficient. In cases where the model with the 3PLN shows better model fit and items differ in their speed sensitivity, using only the hierarchical framework with the 2PLN could lead to unfair testing situations. To be more precise, the ability estimates and the rank order of test takers could heavily depend on the

administered test form, especially for slower test takers. Instead, the hierarchical framework with the 3PLN should be used when calibrating the items, and not only the average testing time but also the sensitivity to speed differences should be balanced across test forms.

Another common alternative for the assembly of fair test forms is the approach of assembling unspeeded test forms. Because in most educational assessments speed is a nuisance parameter that is conceptually not part of the construct being measured, this strategy seems promising. However, the results of this study and results from previous studies (e.g., van der Linden & Xiong, 2013) indicate that this approach might be unfeasible because there are generally large differences in the time that test takers require to respond to all items in an assessment (see Table 2). Assuring that even the slowest test takers can work without time pressure would imply a time limit that is far too generous for fast test takers and problematic both from an economical and from a motivational perspective. Furthermore, the results of this study have important implications for determining the speededness of a test: So far, often experimental methods using different time limits or different numbers of items in the same time limit have been used (e.g., Bridgman, Cline, & Hessinger, 2004; Bridgman, Trapani, & Curley, 2004; Harik et al., 2018). But while for the majority of the test takers more generous time limits might only have a small impact on the demonstrated ability, different time limits can still substantially affect the slowest part of the population. This effect can only be disentangled by explicitly modeling speed. If differences in ability estimation for different time limits are averaged over all test takers or calculated for different ability levels, the degree of speededness of the test for slow test takers could be severely underestimated. Therefore, tests that have been examined using the aforementioned experimental methods could have been falsely classified as unspeeded.

### *Limitations*

There are a number of limitations to this study. First, the real data analysis is based on low-stakes data while implications are mainly relevant for high-stakes assessments. However, similar analyses on (non-educational) high-stakes data have reported similar findings (Fox & Marianti, 2017). In addition, it is not uncommon that pilot studies for item pool calibrations are conducted under low-stakes conditions. Furthermore, this study does not conclude that the hierarchical framework with the 3PLN will always demonstrate better model fit than the model with the 2PLN for item pools of high-stakes assessments. Rather, it argues that the assumption of equal speed sensitivity across items should be tested, just like the assumption of equal factor loadings should be tested in confirmatory factor analysis or structural equation modeling (Brown, 2006).

A second limitation relates to a general limitation of the hierarchical framework, namely, the assumption of stationarity (van der Linden, 2007). The model assumes that given the common distribution of the person and item parameters, residuals between responses and response times are independent. The assumption is, for example, violated if participants substantially speed up or slow down during the test. This could happen in high-stakes assessments with a time limit, if test takers speed up when they feel they are running out of time. However, for test assembly purposes, only item parameters and their relations across items are of interest. If position effects are controlled for (similar to controlling for position effects of ability item parameters estimation, for example, Gonzalez & Rutkowski, 2010), speeding up might only affect the precision of item parameter estimation. Avoiding speeding up seems easiest if items were piloted in low-stakes settings.

A third limitation is that this study deals with a specific violation of the model assumptions of the 2PLN. In the past, assumptions of the hierarchical framework using the 2PLN or 3PLN for response times have been critically reviewed using empirical data analyses (Bolsinova &

Tijmstra, 2018; Klein Entink, van der Linden, & Fox, 2009; Fox & Marianti, 2016; Ranger & Ortner, 2012). Criticism includes violations of the assumption of lognormally distributed response times and the stationarity assumption mentioned above. Although the lognormal distribution has been the standard for modeling response times in educational assessments, future research could explore alternatives to the 2PLN and 3PLN, possibly also embedded in the hierarchical framework.

## *Outlook*

In the past, *automated test assembly* (ATA) procedures have been developed to enable the assembly of multiple test forms from large item pools under various constraints (van der Linden, 2005). These methods are already frequently used in practice (Luecht & Sireci, 2011). To enable the use of the 2PLN in ATA, van der Linden (2011a, 2011b) reparametrized the model. Future research should investigate how the 3PLN can be used best in automated test assembly and whether a similar reparameterization approach might be feasible.

The 3PLN could also be useful to determine the speededness of assessments under various time constraints for different test-taker populations without having to experimentally investigate all possible combinations. This would especially be valuable for determining test accommodations for students with disabilities (Lovett, 2010). Furthermore, while this article focuses on fixed-test forms, the findings can also be applied to computerized adaptive testing or multistage testing. Studies have shown that differential speededness of test forms is an even greater challenge in these settings (van der Linden & Xiong, 2013). Investigating whether the 3PLN could contribute to the fairness of these assessments seems worthwhile as well.

### Declaration of Conflicting Interests

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

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### ORCID iD


Benjamin Becker https://orcid.org/0000-0003-3074-0918

### Supplemental Material

Supplementary material is available for this article online

### Notes

1. Given that ability and speed are distinct constructs, which appears to be a reasonable assumption (e.g., van der Linden, 2009).
2. For a discussion of this issue, see, for example, the work of Tijmstra and Bolsinova (2018).
3. This may be caused by the labeling of the inverse of the residual variance as a discrimination parameter, which is usually a term used for slope parameters. For example, Bertling and Weeks (2018) cite van der Linden (2006) but introduce the model with $\alpha_k$ as a slope parameter instead of the inverse of the residual variance.

4. Note that van der Linden (2006) also includes the inverse of the residual variance $\sigma^2_{\epsilon_k}$ in the joint item parameter distribution. For better comparability with the three-parameter lognormal model (3PLN) in the hierarchical framework, the joint item parameter distribution is slightly modified in this study by assuming a univariate distributed item-specific residual variance, independent from the distribution of the other item parameters (see, for example, also Pohl et al., 2019).
5. In the empirical example, $SD(\phi_k)$ within booklets was around 0.35 and the range across booklets for $\phi_k$ was 0.6.
6. Table 2 contains mean statistics across all replications, whereas standard deviation for the identical statistics across replications can be found in Online Appendix H.

## References


Bertling, M., & Weeks, J. P. (2018). Using response time data to reduce testing time in cognitive tests. *Psychological Assessment*, *30*(3), 328–338.

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 13–38.

Bridgman, B., Cline, F., & Hessinger, J. (2004). Effect of extra time on verbal and quantitative GRE scores. *Applied Measurement in Education*, *17*(1), 25–37.

Bridgman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement*, *41*(4), 291–310.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.

College Board. (2016). *Test specifications for the redesigned SAT* (C. Board, Ed.).

Debelak, R., Gittler, G., & Arendasy, M. (2014). On gender differences in mental rotation processing speed. *Learning and Individual Differences*, *29*, 8–17.

Educational Testing Service. (2020). *Test framework and test development, volume 1* (E. T. Service, Ed.; TOEFL iBT).

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225–245.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer.

Fox, J.-P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 540–553.

Fox, J.-P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement*, *54*(2), 243–262.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–511.

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S. Brooks, A. Gelman, G. L. Jones, & X. L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 163–174). Chapman & Hall/CRC.

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, *13*, 133–164.

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, *39*, 108–119.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet design and parameter recovery in large-scale assessments. In M. von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (*Vol. 3*, pp. 125–156). IEA-ETS Research Institute.

Harik, P., Clauser, B. E., Grabovsky, I., Baldwin, P., Margolis, M. J., Bucak, D., Jodoin, M., Walsh, W., & Haist, S. (2018). A comparison of experimental and observational approaches to assessing the effects of time constraints in a medical licensing examination. *Journal of Educational Measurement*, *55*(2), 308–327.

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21–48.

Klein Entink, R. H., van der Linden, W. J., & Fox, J-. P. (2009). A Box–Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*, 621–640.

Lord, F. M., & Novick, M. R. (1986). *Statistical theories of mental test scores*. Information Age Publishing.

Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research*, *80*(4), 611–638.

Luecht, R. M., & Sireci, S. G. (2011). *A review of models for computer-based testing* (College Board, Ed.; Research Report 2011–12). College Board.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74.

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 technical report* (PISA 2015 technical report). OECD Publishing.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*.

Plummer, M. (2016). *Rjags: Bayesian graphical models using MCMC* (R Package Version 4-6). https://CRAN.R-project.org/package=rjags

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). Coda: Convergence diagnosis and output analysis for MCMC. *R News*, *6*(1), 7–11. https://journal.r-project.org/archive/

Pohl, S., Ulitzsch, E., & von Davier, M. (2019). Using response times to model not-reached items due to time limits. *Psychometrika*, *84*(3), 892–920.

Ranger, J., & Ortner, T. (2012). A latent trait model for response times on tests employing the proportional hazard model. *British Journal of Mathematical and Statistical Psychology*, *65*, 334–349.

Robitzsch, A., Kiefer, T., & Wu, M. (2017). *TAM: Test Analysis Modules* (R Package Version 2.8-21). https://CRAN.R-project.org/package=TAM

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, *42*(2), 193–198.

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, *48*, 37–50.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, *64*(4), 583–639.

Tijmstra, J., & Bolsinova, M. (2018). On the importance of the speed-ability trade-off when dealing with not reached items. *Frontiers in Psychology, 9*, Article 964.

van der Linden, W. J. (2005). *Linear models for optimal test assembly*. Springer.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*(3), 247–272.

van der Linden, W. J. (2011a). Setting time limits on tests. *Applied Psychological Measurement*, *35*(3), 183–199.

van der Linden, W. J. (2011b). Test design and speededness. *Journal of Educational Measurement*, *48*(1), 44–60.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327–347.

van der Linden, W. J., & Xiong, X. (2013). Speededness and adaptive testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 418–438.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.