A Multilevel Mixture IRT Framework for Modeling Response Times as Predictors or Indicators of Response Engagement in IRT Models

Gabriel Nagy1 and Esther Ulitzsch1

Abstract
Disengaged item responses pose a threat to the validity of the results provided by large-scale assessments. Several procedures for identifying disengaged responses on the basis of observed response times have been suggested, and item response theory (IRT) models for response engagement have been proposed. We outline that response time-based procedures for classifying response engagement and IRT models for response engagement are based on common ideas, and we propose the distinction between independent and dependent latent class IRT models. In all IRT models considered, response engagement is represented by an item-level latent class variable, but the models assume that response times either reflect or predict engagement. We summarize existing IRT models that belong to each group and extend them to increase their flexibility. Furthermore, we propose a flexible multilevel mixture IRT framework in which all IRT models can be estimated by means of marginal maximum likelihood. The framework is based on the widespread Mplus software, thereby making the procedure accessible to a broad audience. The procedures are illustrated on the basis of publicly available large-scale data. Our results show that the different IRT models for response engagement provided slightly different adjustments of item parameters of individuals’ proficiency estimates relative to a conventional IRT model.

1Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Corresponding Author:
Gabriel Nagy, Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, Kiel, 24118, Germany.
Email: nagy@ipn.uni-kiel.de
Keywords
item response theory, multilevel mixture modeling, response time, response engagement

Introduction
Large-scale assessments of achievement attempt to assess what individuals know and can do.

In order to provide unbiased results, individuals should ideally answer all items with full engagement (Wise & Smith, 2016). However, this requirement is often not fulfilled, especially in low-stakes assessments that have no consequences for individuals (List et al., 2017). Disengagement is considered a threat to the validity of results, because it may lead to lowered proficiency estimates (Wise, 2015) and biased item parameter estimates (Schnipke & Scrams, 1997).

Since the advent of computerized testing, item response times are most often used to identify disengaged responses. In recent years, a number of item response theory (IRT) models for response engagement that augment item responses by response times have been proposed (Pokropek, 2016; Ulitzsch et al., 2020; Wang & Xu, 2015; Wang et al., 2018). These models have the desirable properties of not requiring explicit classifications of engaged and disengaged responses, of reducing bias in item parameter estimates without discarding responses, and of providing estimates of proficiencies that are adjusted for disengagement. The IRT models build on different assumptions and might therefore give different results. However, in real applications, decisions for one specific model are complicated by the fact that many models have not been implemented in widely accessible statistical software. This limits the applicability of IRT models in practice, although many large-scale programs have made item-level data including response times available (e.g., Goldhammer et al., 2016; OECD, 2017).

The aims of the present article are: (1) to summarize the assumptions underlying the classification of response engagement, (2) to outline how these assumptions are embedded in existing IRT models for response engagement, (3) to present some extensions of these models, and (4) to introduce a flexible multilevel mixture IRT (MM-IRT) framework in which IRT models for response engagement can be specified and estimated.

The MM-IRT framework builds upon the widespread Mplus program (Muthén & Muthén, 1998), thereby making IRT models for response engagement accessible to applied researchers. All models can be estimated by marginal maximum likelihood (MML) with the expectation maximization algorithm (EM; Darrell Bock & Lieberman, 1970), including those models that have previously been estimated only in a Bayesian framework (e.g., Ulitzsch et al., 2020; Wang et al., 2018). Although Bayesian estimation has many advantages, it comes at the price of long estimation times. This renders Bayesian estimation impractical in the case of large sample sizes, which are typical for large-scale assessments. In contrast, MML estimation is mainly
limited by the number of dimensions that need to be numerically integrated, and it can be applied in fairly large samples.

The remainder of this article is structured as follows. In the next section, we outline the rationales underlying traditional procedures for classifying response engagement. We then describe the IRT models for response engagement that have been suggested in the literature, discuss their assumptions, and provide some extensions that make the models more flexible. Following that, we introduce the MM-IRT framework and show how the models presented are specified in this framework. The framework is illustrated with publicly available data taken from the program for the international assessment of adult competencies (http://www.oecd.org/skills/piaac/publicdataandanalysis/).

**Indicators of Response Engagement Based on Item Response Times**

The assessment of response engagement has predominantly focused on rapid guessing (Wise, 2017) as a manifestation of disengagement (for other aspects see van Barneveld et al., 2013). Responses given in a time that falls below a certain threshold are considered to be too fast to reflect solution behavior (SB). Item responses are classified as being engaged or disengaged by means of the SB-index, which has a value of 1 for individual i’s (\( i = 1, 2, \ldots, N \)) response to item j (\( j = 1, 2, \ldots, J \)) if \( i \)'s response time to item j (\( RT_{ij} \)) exceeds the time threshold \( \tau_j \) (\( SB_{ij} = 1 \) if \( RT_{ij} \geq \tau_j \)), and is 0 otherwise (\( SB_{ij} = 0 \) if \( RT_{ij} < \tau_j \)). The SB-index varies within individuals because individuals may be engaged on some items and disengaged on others. It also varies between individuals because individuals could differ in the proportion of engaged responses (Wise & Kong, 2005). Different methods for setting response time thresholds that are based on different rationales have been suggested. Following common approaches for modeling response times (e.g., van der Linden, 2007), we outline the methods while assuming log-transformed response times throughout this article, to which we refer with \( l_{ij} \) (for alternative transformations see De Boeck & Jeon, 2019).

A popular method rests on the assumption of a bimodal response time distribution and looks for the lowest density in the region between the modes (e.g., Wise, 2015). We will refer to this method as the visual inspection of bimodality (VIB) method. As shown in Figure 1, the VIB method assumes a mixture distribution due to engaged and disengaged response processes and aims to identify the lowest response time in engaged responses. The conceptual idea underlying the VIB procedure is visualized in Figure 1 by means of a schematic path diagram. Here, \( SB_{ij} \) denotes an individual \( i \)'s latent engagement state on item j that affects her or his response time \( l_{ij} \). Statistically, SB is represented as a latent class variable and response times as indicators thereof.

A second approach combines response times with item responses (correct vs. incorrect), which are denoted as \( y_{ij} \). This method, referred to here as the VICS method, visually inspects the proportion of correct scores conditional on response times.
Figure 1. Upper panels = examples for procedures for setting response time thresholds based on artificial data.

Lower panels = Graphical depiction of conceptual ideas underlying the procedures for setting response time thresholds.
times. In the VICS method, the threshold is set according to the response time below which the proportion of correct responses is not higher than the chance level of success (Wise & Ma, 2012; for more elaborate versions see Guo et al., 2016; Lee & Jia, 2014). That is, for engaged responses the probability of a correct response is expected to be higher than for disengaged responses. Therefore, as represented by the path diagram in Figure 1, the VICS procedure assumes that item responses are impacted by the latent engagement state, which, in turn, is predicted by response times.

A third method assumes that disengaged responses do not reflect proficiency (i.e., they are uninformative) (Wise, 2019). The method uses response times and correlations of item responses with provisional proficiency estimates (denoted as \( \hat{\theta} \)). Thresholds are based on a visual inspection of plots of the item score’s correlation with \( \hat{\theta} \), so that they refer to the response time at which the correlation exceeds a reference value. We refer to this method as the visual inspection of information (VII) method. An example that uses a reference correlation of 0.20 (Wise, 2019) is given in Figure 1. As shown in the path diagram, the VII method assumes that the latent engagement variable determines the impact of \( \hat{\theta} \) on item responses, whereas response times serve as predictors of the engagement state.

Wise (2019) suggested a method in which the VICS and the VII procedures are combined. Conceptually, in the combined VICS–VII method, success rates and item information are simultaneously considered as outcomes of response engagement, whereas response times serve as predictors of engagement (Figure 1). A different possibility of combining some aspects of the VIB, the VICS, and the VII methods is to consider response times, success rates, and item information as indicators of engagement.

**IRT Models for Response Engagement**

IRT models for response engagement build upon assumptions that parallel some assumptions of the VICS (the proportion of correct disengaged responses corresponds to the chance level) and the VII procedures (disengaged responses are not related to proficiency). The IRT models considered here recur on one dichotomous latent class variable \( C \) that is allowed to vary within individuals, which means that individuals could be engaged on some items and disengaged on others. All models have in common that, for engaged responses \( (C_{ij} = 1) \), they assume an IRT model in which a response to item \( j \), \( y_{ij} \), reflects an individual \( i \)’s standing on the proficiency variable \( \theta \) (for an extension to multidimensional proficiencies see Lu et al., 2020). To this end, different IRT models have been used, including the one-parameter (Ulitzsch et al., 2020), the two-parameter (2PL; Pokropek, 2016), and the three-parameter logistic (Wang & Xu, 2015) IRT model. In the case of disengagement \( (C_{ij} = 0) \) it is assumed that responses do not depend on proficiency and that the probability of a correct response equals the chance level of success.

In the IRT models, the probability of a correct response is structured as

\[
P(y_{ij} = 1) = C_{ij}P(y_{ij} = 1|C_{ij} = 1) + (1 - C_{ij})P(y_{ij} = 1|C_{ij} = 0),
\]

(1)
where $P(v_{ij} = 1|C_{ij} = 1)$, the probability of a correct response in the engaged state, is parameterized according to an IRT model, whereas $P(v_{ij} = 1|C_{ij} = 0)$ is modeled by a single item parameter to reflect the chance probability of success.

When the latent class variable is replaced with the SB-index, the structure corresponds to the effort-moderated IRT model of Wise and DeMars (2006), which estimates the IRT parameters and the individuals’ proficiency levels solely on the basis of responses that, through the use of threshold procedures, have been classified as engaged. The IRT models for response engagement enable researchers to bypass the cumbersome procedures of classifying the engagement status of each item response (Figure 1) and make it possible to deal with responses for which the engagement status cannot be univocally classified.

Similar to the procedures for setting response time thresholds, different kinds of IRT models for response engagement have been suggested that treat response times either as indicators or as predictors of the engagement states. The most commonly used IRT models (Ulitzsch et al., 2020; Wang & Xu, 2015; Wang et al., 2018) combine Schnipke and Scrams’s (1997) mixture model for response times with an IRT part and consider both response times and item responses as being influenced by the class variable. Therefore, we will refer to this group of models as independent latent class IRT (ILC-IRT) models. In a second type of mixture IRT models, the within-individual latent class variable is treated as a dependent variable that is predicted by response times (e.g., Pokropek, 2016). We will refer to this group of models as dependent latent class IRT (DLC-IRT) models.

**DLC-IRT Models**

DLC-IRT models where response times are used as predictors of response engagement have not often been used in practice. An exception is a model suggested by Pokropek (2016). In the following, we review this model and suggest some extensions to increase its flexibility.

**DLC-IRT Models with Single-Level Relationships.** Pokropek (2016) suggested a DLC-IRT model in which engagement states are predicted by within-individual level response times. Because relationships between response times, response engagement, and item responses are located solely on one level in Pokropek’s (2016) model, we refer to this type of model as a DLC-IRT model with single-level relationships (DLC-SL-IRT).

In contrast to Pokropek’s (2016)’s original model, we use log-transformed predictors, and we start with a slightly different but statistically equivalent parameterization. In this setup, the probability of an engaged response is expressed as

$$P(C_{ij} = 1) = \frac{\exp[\gamma(l_{ij} - \tau)]}{1 + \exp[\gamma(l_{ij} - \tau)]},$$

where $\tau$ is a response time threshold that is common to all items and individuals so that the probability of engaged behavior is 0.5 for log response times equal to $\tau$. In
addition, $\gamma (\gamma > 0)$ is a logistic regression weight that can be understood as a process discrimination parameter. The higher the values of $\gamma$, the better response times close to $t$ discriminate between engaged and disengaged responses.

Pokropek’s (2016) model implies that within an individual $i$, a response to item $j$ is more likely to be engaged than a response to another item $k$ when $l_{ij} > l_{ik}$. Such comparisons are problematic because, typically, items require different amounts of time, and the prevalence of engaged responses differs between items (e.g., Wise, 2019). These differences can be accounted for by making the thresholds item-specific, so that

$$P(C_{ij} = 1) = \frac{\exp[\gamma (l_{ij} - \tau_j)]}{1 + \exp[\gamma (l_{ij} - \tau_j)]}. \quad (3)$$

Item-specific thresholds $\tau_j$ imply that, within an individual $i$, a response to item $j$ is more likely to be engaged than a response to item $k$ when $(l_{ij} - \tau_j) > (l_{ik} - \tau_k)$.

The DLC-SL-IRT model is sketched as a schematic path diagram in Figure 2. Response engagement is specified as a within-individual latent class variable $C$. The arrow leading from $C$ to the item response means that success probabilities differ between engaged and disengaged responses. The proficiency variable $\theta$ is located on the between-individual level, and its impact on the observed item response is moderated by $C$ (i.e., disengaged item responses do not reflect proficiency). In the DLC-SL-IRT model, $C$ is predicted by response times. The gray-shaded latent variable stands for a node variable that is a function of the item-specific response time and thresholds $(l_{ij} - \tau_j)$. This variable is related to $C$ by the process discrimination parameter $\gamma$.

**DLC-IRT Models with Two-Level Relationships.** The DLC-SL-IRT model implies that for each individual, the proportion of engaged responses can be fully explained by their item-specific log response times. This is because the model does not include between-individual differences in the latent class variable. However, Pokropek (2016) (see also Rios et al., 2017) noted that when between-individual differences that are conditional on response times exist in the proportion of engaged responses and these are correlated with proficiency, proficiency estimates are likely to be biased. From a conceptual point of view, such relationships are plausible. More proficient individuals have been found to need less time to complete an item conscientiously (e.g., Goldhammer et al., 2013), which means that, for such individuals, a response to item $j$ that is given slightly below the response time threshold $\tau_j$ could still be engaged. Conversely, for individuals of lower proficiency, response times that fall slightly above the response time threshold might be more likely to reflect disengaged behavior. Therefore, response time thresholds could be person-specific and correlated with proficiency. Based on these considerations, we suggest an extension of the DLC-SL-IRT model that includes between-individual differences in response time thresholds. As these components extend the model by between-individual level
aspects, we refer to the resulting model as the DLC-IRT model with two-level relationships (DLC-TL-IRT).

In the DLC-TL-IRT model, we represent the impact of response times as

\[
PC_{ij} = \frac{1}{C_0/C_1} \exp\left(\frac{\gamma_{ij}^{\mathrm{li}}}{C_0}\frac{t_j}{C_0/C_1/C_2/C_3/C_8/C_9}\right) + \frac{C_0/C_1}{C_0/C_1/C_2/C_3/C_8/C_9},
\]

where \(t_j\) is the response time threshold for item \(j\), and \(\zeta_i\) is a normally distributed individual-specific threshold disturbance \(\zeta \sim N(0, \sigma_\zeta^2)\) that accounts for between-individual differences in response time thresholds. Individuals with higher values on \(\zeta\) have higher than average response thresholds, which means that their responses are

\[P(C_{ij} = 1) = \frac{\exp\left(\gamma [l_j - (\tau_j + \zeta_i)]\right)}{1 + \exp\left(\gamma [l_j - (\tau_j + \zeta_i)]\right)},\]

\(4\)

Figure 2. Schematic representation of different IRT models for response engagement. The dotted horizontal line separates the within-individual (lower part) from the between-individual (upper part) level of analysis. Gray filled circles represent node variables that are functions of other observed and latent variables. Triangles are used to indicate intercepts and thresholds.
more likely to be disengaged than the responses of individuals who respond in the same time but have lower values on $\zeta$.

Figure 2 includes a graphical depiction of the DLC-TL-IRT model. The model extends the DLC-SL-IRT model by including the between-individual level threshold disturbance $\zeta$. The node variable evaluates to $l_{ij} - (T_j + \zeta_j)$, and the corresponding term is related to $C$ via $\gamma$. The proficiency variable $\theta$ is allowed to correlate with $\zeta$. A negative correlation between $\theta$ and $\zeta$ indicates that more proficient individuals are more likely to provide thorough responses even when responses are relatively fast.

Posterior Engagement Probabilities in DLC-IRT Models. Equations (3) and (4) refer to the probability of an engaged response that are expected on the basis of response times and, therefore, do not provide a full insight into the probability that a response was actually engaged. To better understand the classification of responses in DLC-IRT models, it is instructive to consider the posterior probability of individual $i$’s response to item $j$ being engaged, which we refer to as $p_{ij}$. The posterior engagement probabilities depend on the response times and the observed item responses and can be expressed as

$$p_{ij} = \frac{P(y_{ij}|C_{ij} = 1)P(C_{ij} = 1)}{P(y_{ij}|C_{ij} = 1)P(C_{ij} = 1) + P(y_{ij}|C_{ij} = 0)P(C_{ij} = 0)},$$

where $P(y_{ij}|C_{ij} = 1)$ and $P(y_{ij}|C_{ij} = 0)$ refer to the probability of the observed item response in the engaged and the disengaged response state, respectively. These probabilities depend on the item responses (correct vs. wrong). For example, $P(y_{ij} = 1|C_{ij} = 0)$ refers to the chance level of success of item $j$, whereas $P(y_{ij} = 0|C_{ij} = 0) = 1 - P(y_{ij} = 1|C_{ij} = 0)$ reflects $j$’s chance level of failure. Typically, the chance level of success is low or even zero (e.g., in the case of open responses), which means that the chance level of failure is close to one or even one. In such typical situations, the posterior engagement probabilities of wrong responses will mainly reflect $P(C_{ij} = 1)$, which is monotonically related to $l_{ij}$ (Equations 3 and 4). In contrast, correct responses are likely to receive high posterior engagement probabilities even when they are given in a short amount of time.

Interim Summary of DLC-IRT Models. In DLC-IRT models, the meaning of the latent class variable is defined by the item responses, so that disengaged responses have a solution probability that corresponds to the chance level of success and are not related to proficiency. Response times are used as predictors of latent class membership, which means that response times are the main source of information used to separate engaged from disengaged responses. In the typical case of a low chance level of success, the posterior engagement probabilities in DLC-IRT models reduce the contribution of wrong responses coupled with fast response times in the estimation of IRT item parameters and proficiencies.
Therefore, DLC-IRT models are very similar to the combined ICS-VII procedure for setting response time thresholds (Figure 1). Of the two DLC-IRT models, the DLC-SL-IRT model is most similar to the VICS-VII procedure because it assumes that the response times contain all information necessary for predicting the engagement status of item responses. The DLC-TL-IRT model relaxes some restrictive assumptions, namely, that response times unequivocally predict the engagement status of each response (random threshold) and that conditional on response times engagement is not related to proficiency (correlation between random threshold and proficiency).

**ILC-IRT Models**

In ILC-IRT models, response times are considered to be indicators of response engagement. The basis of the most common type of ILC-IRT model was laid out by Schnipke and Scrams (1997). Their approach was extended by Wang and Xu (2015), Wang et al. (2018), and Ulitzsch et al. (2020), who incorporated item responses into the models by using the structure shown in Equation (1) and imposing a similar structure on response times:

\[
\frac{\gamma_{ij}}{1 + \frac{1}{C_0/C_1} + \frac{1}{C_0}} = \frac{1}{C_0/C_1} + \frac{1}{C_0} \frac{C_{ij}}{C_0/C_1} \quad (6)
\]

ILC-IRT models include an additional latent variable \( \xi \) that underlies response times. Typically, \( \xi \) is specified to reflect individual differences in working speed (higher values on \( \xi \) are associated with shorter response times; see Klein Entink et al., 2009; van der Linden, 2006). Here, we use a statistically equivalent parameterization in which \( \xi \) reflects individuals’ typical time expenditure (higher values on \( \xi \) are associated with longer response times). We assume a congeneric measurement model that corresponds to

\[
(I_{ij} \mid C_{ij} = 1) = \nu_j + \lambda_j \xi_i + \epsilon_{ij} \quad (7)
\]

for engaged responses and to

\[
(I_{ij} \mid C_{ij} = 0) = \bar{\nu}_j + \bar{\lambda}_j \xi_i + \bar{\epsilon}_{ij} \quad (8)
\]

for disengaged responses. The tilde over the intercepts (\( \nu_j \)), loadings (\( \lambda_j \)), and residual variables (\( \epsilon_{ij} \)) indicates that the intercepts, loadings, and residual variances (\( \sigma^2_{\epsilon_{ij}} \) and \( \sigma^2_{\bar{\epsilon}_{ij}} \)) can differ between engaged and disengaged responses.

ILC-IRT models differ in how the class probabilities are expressed. Ulitzsch et al. (2020) provided the most comprehensive parameterization, which corresponds to

\[
P(C_{ij} = 1) = \frac{\exp(\eta_i - \kappa_j)}{1 + \exp(\eta_i - \kappa_j)}, \quad (9)
\]
where $\eta$ is a normally distributed person characteristic $[\eta \sim \mathcal{N}(0, \sigma^2_\eta)]$, and $\kappa_j$ is a parameter specific to item $j$. The $\eta$-variable captures the individuals’ propensities of being engaged, whereas the $\kappa$-parameters govern the degree to which the items evoke disengagement. The discrepancy between an individual $i$’s engagement propensity $\eta_i$ and an item $j$’s engagement difficulty $\kappa_j$ determines the probability of engagement of individual $i$ on item $j$. Equation (9) implies unidimensional engagement propensities and conditional independence of item-specific engagement states.

ILC-IRT models that include the $\eta$-variable make it possible to examine the relationships of engagement propensity $\eta$ with $\theta$ and could be beneficial for reducing biases in proficiency estimates (Pokropek, 2016; Rios et al., 2017). Wang and Xu (2015) and Wang et al. (2018) did not include an individual difference variable $\eta$ in their models. Wang and Xu (2015) incorporated only one fixed effect for modeling $P(C_{ij} = 1)$, whereas Wang et al. (2018) included item effects ($\kappa_j$).

**ILC-IRT Models With Random Effects of Response Engagement.** Up until today, applications of ILC-IRT models have employed a constrained version of the structure provided in Equations (7) and (8). For engaged responses (Equation 7), the $\lambda$-parameters are constrained to be equal and the residual variances to be the same for all items $[\tilde{e}_{ij} \sim \mathcal{N}(0, \sigma^2_{\tilde{e}})]$, whereas for disengaged responses (Equation 7), the loadings are fixed to zero ($\lambda_j = 0$ for all $j = 1, 2, \ldots, J$), and the intercepts $\tilde{\nu}_j$ and residuals are set to be equal across all items $[\tilde{\nu}_j = \tilde{\nu}$ and $\tilde{e}_{ij} \sim \mathcal{N}(0, \sigma^2_{\tilde{e}})$ for all $j = 1, 2, \ldots, J$]. The constraints imposed on disengaged responses mean that their distribution is assumed to be the same for all items and individuals $[(l_j | C_j = 0) \sim \mathcal{N}(\tilde{\nu}, \sigma^2_\tilde{e})]$. This specification was introduced by Schnipke and Scrams (1997), who argued that faster response times are a result of examinees skimming the items in search of cues without thoroughly processing the problem statement. As such, item and person characteristics are expected to have negligible effects on disengaged response times.

The restrictions imposed on disengaged response times offer a substantive and a statistical interpretation of $\xi$. The substantive interpretation is that only engaged responses allow the assessment of the typical time expenditure. According to the statistical interpretation, $\xi$ is an individual-specific random effect of response engagement on response times. Therefore, we refer to these kinds of models as ILC-RE-IRT models.

To outline the random effect interpretation, we express the effect of $C_j$ on $l_j$ by inserting Equations (7) and (8) into Equation (6). After some rearrangement, that gives

$$l_{ij} = \tilde{\nu}_j + C_{ij}(\delta_j + \lambda_j \xi_j) + C_{ij}\tilde{e}_{ij} + (1 - C_{ij})\tilde{e}_{ij}, \quad (10)$$

where $\delta_j$ corresponds to $\delta_j = \nu_j - \tilde{\nu}_j$. In Equation (10), the term in the first parentheses, $\delta_j + \lambda_j \xi_j$, refers to the effect of $C_j$ on $l_j$, $\tilde{\nu}_j$ to a regression intercept, and the remaining terms capture the possibly heteroscedastic residual variance of $l_j$. The effect of the latent class variable comprises two parts: (1) an average difference in response times $\delta_j$ and (2) systematic individual deviations thereof $\lambda_j \xi_j$. Therefore, in
ILC-RE-IRT models, the $\xi$-variable represents individual differences in the separation of engaged and disengaged response times.

Figure 2 includes a schematic path diagram of the ILC-RE-IRT model. The latent class variable reflects the between-individual component $\eta$ and includes an item-specific threshold $\kappa_j$ that governs the engagement difficulty of item $j$. The engagement state affects the log response time to item $j$ with an average effect $\delta_j$ (Equation 10). The between-individual level variable $\xi$ captures individual differences in this effect because its loadings are zero when $C_j = 0$ and $\lambda_j$ when $C_j = 1$. The latent variables are allowed to correlate. The correlation between $\eta$ and $\xi$ indicates whether higher engagement propensities are related to higher or lower effects of response engagement on response times.

**ILC-IRT Models with Random Intercepts of Response Engagement.** The assumption that disengaged response times do not reflect typical time expenditure could be questioned when items elicit time-consuming actions even when responses are disengaged. For example, many items contain long reading passages, figures, and tables that require examinees to spend some time even if they examine the material only superficially. In addition, items that have an open response format require examinees to write down an answer, which is more time-consuming than randomly crossing one response option for a multiple choice item (Wise, 2019). In such cases, it is reasonable to assume that both engaged and disengaged response times depend on individuals’ typical time expenditure.

This assumption can be incorporated into ILC-IRT models by allowing the loadings $\lambda_j$ in Equation (8) to differ from zero. We suggest imposing the invariance constraint $\lambda_j = \lambda_j$ (Equations 7 and 8), which leads to a model with a clear interpretation. Based on this constraint, log response times can be expressed as an effect of the latent class variable as

$$l_{ij} = \tilde{\nu}_j + \lambda_j \xi_i + C_{ij} \delta_j + C_{ij} \epsilon_{ij} + (1 - C_{ij}) \tilde{\epsilon}_{ij}.$$  \hspace{1cm} (11)

Here, the $\delta$-parameters ($\delta_j = \nu_j - \tilde{\nu}_j$) represent the effect of the latent class variable that does not vary between individuals and the sum $\tilde{\nu}_j + \lambda_j \xi_i$ can be interpreted as a random intercept, with $\tilde{\nu}_j$ denoting the average intercept and $\lambda_j \xi_i$ the systematic individual deviations thereof. Similar to Equation (10), the remaining terms are used to capture the possibly heteroscedastic residual variances. Therefore, we refer to the model based on Equation (11) as the ILC-IRT model with a random intercept of response times (ILC-RI-IRT). In ILC-RI-IRT models, the $\xi$-variable represents overall individual differences in response times, so that engaged and disengaged response times are defined relative to $\xi$.

The structure of ILC-RI-IRT models that also include $\eta$ is shown in Figure 2 with a path diagram. The only difference to the ILC-RE-IRT model is that the response times’ loadings are not affected by the latent class variable. The engagement state of an individual response is affected by $\eta$, which means that $\eta$ affects the log response times via the latent class variable. Therefore, the correlation of $\eta$ with $\xi$ indicates
whether between-individual differences in engagement propensity are related to indi-
vidual differences in time expenditure over and above the effects of engagement
states. For example, a positive correlation between \( \eta \) and \( \xi \) indicates that individuals
with high engagement propensities are expected to spend more time in either engage-
ment state.

**Posterior Engagement Probabilities in ILC-IRT Models.** In ILC-IRT models the posterior
engagement probabilities \( p_{ij} \) can be represented as

\[
p_{ij} = \frac{f(I_{ij}|C_{ij} = 1)P(y_{ij}|C_{ij} = 1)P(C_{ij} = 1)}{f(I_{ij}|C_{ij} = 1)P(y_{ij}|C_{ij} = 1)P(C_{ij} = 1) + f(I_{ij}|C_{ij} = 0)P(y_{ij}|C_{ij} = 0)P(C_{ij} = 0)},
\]

where \( P(y_{ij}|C_{ij} = 1) \) and \( P(y_{ij}|C_{ij} = 0) \) are as defined in the context of Equation (5)
and \( P(C_{ij} = 1) \) and \( P(C_{ij} = 0) = 1 - P(C_{ij} = 1) \) are as specified in Equation (9). In
ILC-IRT models the posterior engagement probabilities account for response times
by referring to their normal probability density function (PDF) values, defined with
respect to the distribution of engaged \( f(I_{ij}|C_{ij} = 1) \) and disengaged response times
\( f(I_{ij}|C_{ij} = 0) \). The PDF values indicate how typical an observation \( l_{ij} \) is for the
engaged or disengaged state, respectively. Similar to the posterior probabilities in
DLC-IRT models, in the face of low chance levels of success, correct responses are
likely to receive high posterior engagement probabilities, which means that the
response times’ PDF values have a stronger impact on the posterior probabilities of
wrong responses.

Typically, the means of disengaged responses are lower than the means of engaged
responses. Therefore, long response times might be expected to be more typical for
the engaged than the disengaged state (i.e., a higher PDF in the engaged than in the
disengaged response time distribution). However, when the residual variance \( \sigma^2_{\tilde{e}} \) is
much larger than \( \sigma^2_{\tilde{y}} \), the PDF value of a long response time derived with respect to
the disengaged state could be larger than the PDF value in the engaged state.\(^2\) In this
case, ILC-IRT models could indicate that short and very long response times are both
more typical of the disengaged than the engaged state, which means that they both
receive low posterior engagement probabilities. We observed that, in applications of
ILC-IRT models, estimates of \( \sigma^2_{\tilde{e}} \) tended to be much larger than estimates of \( \sigma^2_{\tilde{y}} \) (e.g.,
Schnipke & Scrams, 1997; Ulitzsch et al., 2020). For this reason, the specification of
ILC-IRT models should be carefully thought to prevent biased estimates of residual
variances. For example, the assumption that a single measurement intercept applies
to all disengaged response times should be reconsidered, because erroneously fixing
intercepts to be equal could artificially inflate the estimates of \( \sigma^2_{\tilde{e}} \).

**Interim Summary of ILC-IRT Models.** In ILC-IRT models, responses and response times
jointly define the meaning of the latent class variable. Engaged item responses are
assumed to be indicative of proficiency and to be accompanied by a specific
distribution of response times. In contrast, disengaged responses are assumed to not reflect proficiency and to be accompanied by a different distribution of response times.

We presented two alternative ways of specifying ILC-IRT models for response engagement. The commonly used ILC-RE-IRT models (Ulitzsch et al., 2020; Wang & Xu, 2015; Wang et al., 2018) specify the relationships of the latent class variable with response times by means of a random effect regression, whereas ILC-RI-IRT models replace the random effect with a random intercept. These two specifications imply a different definition of typical time expenditure, so that $\xi$ is defined solely on the basis of engaged responses in ILC-RE-IRT models and on the basis of both kinds of responses in ILC-RI-IRT models. In ILC-RE-IRT models, $\xi$ represents individual differences in the separation of engaged response times from an unstructured distribution that is assumed to be the same for all individuals. In ILC-RI-IRT models, $\xi$ represents overall individual differences in response times, so that engaged and disengaged response time distributions are defined relative to $\xi$.

ILC-IRT models are conceptually related to the VIB method (Wise, 2015), but this connection is not as close as the one between DLC-IRT models and the VICS-VII method. Clearly, continuous response time variables carry more information than categorical item responses, so that latent class membership is more strongly determined by response times than by item responses (Wang & Xu, 2015). However, ILC-IRT models also account for item responses, whereas the VIB method does not. In addition, the VIB procedure aims to separate fast disengaged responses from engaged responses on the basis of a response time threshold (monotonic relationship), whereas ILC-IRT IRT models allow for more complex relationships between response times and engagement.

**DLC-IRT versus ILC-IRT Models**

DLC-IRT and ILC-IRT models provide two different ways to include response times in IRT models for response engagement. Table 1 summarizes how response times are used in the different models considered in this article. In ILC-IRT models, response times are specified as reflections of response engagement, whereas in DLC-IRT models, response engagement is predicted by the individuals’ response times. Therefore, DLC-IRT and ILC-IRT models differ in how dependencies between item responses and response times are modeled. DLC-IRT models consider the distribution of item responses conditional on response times, which means that they take response times as given, without incorporating specific assumptions about their distribution. In contrast, in ILC-IRT models, the joint distribution of item responses and response times is modeled on the basis of assumptions about the structure and the distribution of response engagement and response times. These assumptions make it possible to estimate complex relationships between response engagement and response times that could be non-monotonic (Equation 12) and could differ in strengths between items (Equations 10 and 11). In contrast, in DLC-IRT models, item-specific response times
are assumed to have monotonic relationships of equal strength with response engagement. Given the fundamental differences between ILC-IRT and DLC-IRT models, decisions in favor of one type of model should be based on substantive considerations. Decisions cannot rely on goodness-of-fit indices based on the model likelihood (e.g., Akaike information criterion and Bayesian information criterion) because DLC-IRT and ILC-IRT models consider different aspects of the data. However, such indices can be used to decide between variants of either type of model.

The most important aspect to consider when deciding on a model type is the assumptions about the role of response times and the definition of response engagement. DLC-IRT models take a perspective that is typical for the rapid guessing literature, according to which response engagement can be expressed as a monotonic

Table 1. Summary of the Key Aspects of the Treatment of Item Response Times in Dependent Latent Class (DLC-IRT) and Independent Latent Class (ILC-IRT) Models for Response Engagement.

| Dependent latent class IRT models | DLC-SL-IRT | DLC-TL-IRT |
|-----------------------------------|------------|------------|
| Common aspects                    | Response times predict engagement. | Estimates response times at which an equal probability of engaged and disengaged responses is expected (response time thresholds). |
|                                   | Conditional on item responses, longer response times are monotonically related to higher probabilities of engaged responses. | Response time thresholds could be item-specific. |

| Independent latent class IRT models | ILC-RE-IRT | ILC-RI-IRT |
|-------------------------------------|------------|------------|
| Common aspects                      | Response times reflect engagement. | Response times reflect engagement. |
|                                    | Response times contribute to the classification of engagement by their relative typicality for the engaged and disengaged state. | Response times contribute to the classification of engagement by their relative typicality for the engaged and disengaged state. |
|                                    | Response times can be non-monotonically related to the probability of engaged responses. | Disengaged response times are uncorrelated with each other and have a distribution that is the same for all individuals. |
|                                    | DLC-SL-IRT = dependent latent class IRT model with single-level relationships of response times; DLC-TL-IRT = dependent latent class IRT model with two-level relationships of response times; ILC-RE-IRT = Independent latent class IRT model with a random effect of the latent class variable on response times; ILC-RI-IRT = Independent latent class IRT model with a random intercept of response times. |

Note.
function of response times (the longer the response time, the more likely it is that the response is engaged; Wise, 2019). Furthermore, in this stream of research, it is not common to make structural and distributional assumptions about response times. Therefore, researchers favoring the perspective pursued in research on rapid guessing might find DLC-IRT models to be most attractive. Indeed, DLC-IRT models provide many benefits compared to the typically time-consuming and cumbersome two-step procedures employed in this field (threshold-based classification of response engagement followed by the IRT analyses of responses classified as engaged). DLC-IRT models are one-step procedures that are easy to apply. Here, response time thresholds are estimated within the model, and the model-implied posterior engagement probabilities are accounted for in the estimation of IRT item parameters and individual proficiency levels. Furthermore, DLC-IRT models have the appealing feature that they can examine individual differences in response time thresholds (DLC-TL-IRT), whereas this is not possible on the basis of traditional methods.

ILC-IRT models contain the key ingredients of van der Linden’s (2007) speed-accuracy IRT model, which is frequently used in psychometrically inspired research. This model is based on the assumption that individuals have their own typical working speed (or typical time expenditure), which they maintain over the course of a test. One key research question employed on the basis of this model refers to the relationships between proficiency and typical time expenditure (e.g., Goldhammer et al., 2013). Disengaged item responses violate the constant time expenditure assumption, which means that ILC-IRT models can be seen as extensions of the speed-accuracy IRT model that account for such violations. As such, ILC-IRT models are a suitable choice when researchers are interested in the relationships between proficiency and typical time expenditure and wish to account for disengaged responses defined by atypical response times. In this vein, the two variants of ILC-IRT models, is whether the log response times’ loadings interact (ILC-RE-IRT) or do not interact (ILC-RI-IRT) with the latent class variable, might prove useful in deriving accurate representations of individuals’ typical time expenditure and engagement states.

The MM-IRT for Response Engagement

In this section, the MM-IRT framework, which allows the previously presented IRT models to be estimated with Mplus (Muthén & Muthén, 2017), is introduced. We first describe how the input data are structured and then move to the specification and estimation of DLC-IRT and ILC-IRT models. The section ends with an illustration of the framework.

Data Setup

IRT models can be formulated and estimated as multilevel models with item responses organized in a long format (e.g., Van Den Noortgate et al., 2003). This setup can be extended to MM-IRT models by including an item-level latent class
variable (e.g., Asparouhov & Muthén, 2008; Pokropek, 2016). However, in our experience, the estimation of MM-IRT models on the basis of a long data format is computationally very demanding because the models require the estimation of a large number of interactions. As a solution to this problem, we propose a rearrangement of the input data that greatly speeds up the estimation times when employing MML estimation.

In our setup, the item responses are organized in a diagonal format. Each individual’s item responses are represented in a matrix $Y_i$ of order $J \times J$. The elements of $Y_i$ are denoted as $y_{ijk}$, with missing responses in the off-diagonals ($j \neq k$). Individuals’ arrays of item responses are augmented by a $J \times J$ matrix $D_i$ that contains item indicators $d_{ijk}$. In $D_i$, off-diagonal entries are zero, and diagonal entries are one. Finally, the organization of response times differs between DLC-IRT and ILC-IRT models. In DLC-IRT models, response times are entered in a long format, which means that they are organized in a $J$-valued vector $t_i$ with entries $l_{ij}$. In the case of ILC-IRT models, response times are organized in a $J \times J$ diagonal matrix $T_i$. In this case, entries are denoted as $l_{ijk}$, and only diagonal entries ($j = k$) contain nonmissing values. Table 2 provides an example of a data array for one individual, where response times were entered twice to exemplify their organization in the case of DLC-IRT and ILC-IRT models.

**Specification of IRT Models for Response Engagement in the MM-IRT Framework**

**Specification of the IRT Part for Item Responses.** We first outline the specification of the IRT part that is used to model item responses in all models. In our MM-IRT framework, item responses are represented by a 2PL model that provides a good compromise between flexibility and parsimony and is well suited for the open responses considered in this article.

To include the 2PL part in the MM-IRT framework, we specify the item responses to be located on the within-individual level and the proficiency variable to be located on the between-individual level. For engaged item responses, the 2PL structure is parameterized as

$$P(y_{ijk} = 1 | C_{ij} = 1) = \frac{\exp(\alpha_k w_{ij} - \beta_k)}{1 + \exp(\alpha_k w_{ij} - \beta_k)}, \quad (13)$$

where $\alpha_k$ and $\beta_k$ are the discrimination and threshold parameters of item $j = k$. The term $w_{ij}$ stands for the individual $i$’s scores on a node variable that is fully determined by a $J \times 1$ unit vector $u_i$ with elements $u_{ij} = 1$:

$$w_{ij} = \theta_i u_{ij}, \quad (14)$$

which means that the proficiency variable $\theta$ is represented by a random effect located at the between-individual level. For model identification, we specify $\theta$ to follow a
Table 2. Example of a Data Array for One Individual in the Diagonal Format. Variables in the Columns are Indexed by Item \((k = 1, 2, \ldots, K \text{ with } K = J)\). The Example includes Two Alternative Organizations of Log Response Times as Used in DLC-IRT and ILC-IRT Models.

| Item | Item responses | Item indicators | Log response times |
|------|----------------|-----------------|-------------------|
|      | DLC-IRT and ILC-IRT | DLC-IRT and ILC-IRT | DLC-IRT | ILC-IRT |
| Item | \(y_1\) | \(y_2\) | \(\ldots\) | \(y_{K-1}\) | \(y_K\) | \(d_1\) | \(d_2\) | \(\ldots\) | \(d_{K-1}\) | \(d_K\) | \(l\) | \(l_1\) | \(l_2\) | \(\ldots\) | \(l_{K-1}\) | \(l_K\) |
| 1    | 1     | 1    | \(\ldots\) | 1     | 1    | 0     | 0     | 0     | 3.2 | 3.2 | 3.2 | \(\ldots\) | 3.2 | 3.2 | \(\ldots\) | 3.2 | 3.2 | 3.2 | \(\ldots\) | 3.2 | 3.2 | 3.2 |
| 2    | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | 2.7 | 2.7 | 2.7 | \(\ldots\) | 2.7 | 2.7 | \(\ldots\) | 2.7 | 2.7 | 2.7 | \(\ldots\) | 2.7 | 2.7 | 2.7 |
| \(J-1\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | 1.9 | 1.9 | 1.9 | \(\ldots\) | 1.9 | 1.9 | \(\ldots\) | 1.9 | 1.9 | 1.9 | \(\ldots\) | 1.9 | 1.9 | 1.9 |
| \(J\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | \(\ldots\) | 2.5 | 2.5 | 2.5 | \(\ldots\) | 2.5 | 2.5 | \(\ldots\) | 2.5 | 2.5 | 2.5 | \(\ldots\) | 2.5 | 2.5 | 2.5 |

Note: “–99” denotes missing values.
standard normal distribution \([\theta \sim N(0, 1)]\). Because \(u_i\) is a unit vector, \(w_{ij}\) in Equation (13) can be substituted with \(\theta_i\), which leads to the typical notation of the 2PL model.

For disengaged responses, the success probability does not depend on proficiency, and the probability of a correct response corresponds to the chance level of success. In this case, Equation (13) is altered to

\[
P(y_{ijk} = 1| C_{ij} = 0) = \frac{\exp(\tilde{\alpha}_k w_{ij} - \tilde{\beta}_k)}{1 + \exp(\tilde{\alpha}_k w_{ij} - \beta_k)},
\]

with \(\tilde{\alpha}_k = 0\) for all \(j = k = 1, 2, \ldots, K\) and where \(\tilde{\beta}_k\) governs the chance level of success of item \(j = k\). The \(\tilde{\beta}\)-parameter might be specified as being item-specific or invariant across items and can be freely estimated or fixed to meaningful values. For example, in the case of open responses, the chance level of success is essentially zero, which means that the \(\tilde{\beta}\)-parameters might be fixed to a large value (e.g., 15).

**Specification and Estimation of DLC-IRT Models.** The estimation of the variants of DLC-IRT models recurs on alternative but statistically equivalent parameterizations of Equations (3) and (4). We begin with the more complex DLC-TL-IRT model because the DLC-SL-IRT model is a constrained version thereof. In the DLC-TL-IRT model the latent class variable is specified to depend on item response times that are organized in a long format (Table 2) and on the \(K = J\) item indicators via a logistic regression:

\[
P(C_{ij} = 1) = \frac{\exp\left(\gamma_0 l_{ij} + \sum_{k=1}^{K} \gamma_k d_{ijk} + \psi_i\right)}{1 + \exp\left(\gamma_0 l_{ij} + \sum_{k=1}^{K} \gamma_k d_{ijk} + \psi_i\right)},
\]

where \(\gamma_0\) corresponds to the process discrimination parameter \(\gamma\) in Equation (3). In contrast to Equations (3) and (4), the \(\gamma_k\)-parameters can be interpreted as item-specific logistic intercepts. The term \(\psi_i\) is a random disturbance \([\psi \sim N(0, \sigma_\psi^2)]\) that represents a rescaled counterpart of the DLC-TL-IRT model’s threshold disturbance \(\zeta_i\) in Equation (4). The DLC-SL-IRT model can be derived from Equation (16) by omitting \(\psi\) from the model. The parameters provided by the MM-IRT specification can be converted into the parameters of Equations (3) and (4) as follows. The process discrimination parameter corresponds to \(\gamma = \gamma_0\), the response time thresholds are derived as \(\tau_j = -\gamma_k/\gamma_0\) for \(j = k\), the variance of random disturbance \(\zeta\) equals \(\sigma_\zeta^2 = \sigma_\psi^2/\gamma_0^2\), and the correlation of \(\theta\) with \(\zeta\) is given by \(\rho_{\theta \zeta} = -\sigma_{\theta \psi}/\sigma_\psi\). The standard errors of the converted parameters are provided by the delta method.

The DLC-IRT models can be estimated by MML through the EM algorithm. MML estimation follows from the procedures outlined by Vermunt (2003). The log likelihood to be maximized is given by
\[
\log L = \sum_{i=1}^{N} \log L_i(Y_i|t_i; \phi), \quad (17)
\]

where \( \phi \) represents the vector of model parameters.

As only the diagonal entries \((j = k)\) of \( Y_i \) include information, Equation (17) is equivalent to the log-likelihood expression in which item responses are represented as a \( J \times 1 \) vector \( y_i \). Therefore, the probability density of the data remains the same regardless of how the data are organized. The probability density of the DLT-TL-IRT model can be written as

\[
L_i(Y_i|t_i; \phi) = \int \left[ \prod_{j=1}^{J} P(y_{ij}|t_i, \psi, \theta; \phi) \right] g(\psi, \theta; \phi) d\psi d\theta, \quad (18)
\]

where individual responses are indexed as \( y_{ij} \) to highlight that only diagonal entries in \( Y_i \) \((j = k)\) contribute to the estimation. When the DLC-SL-IRT model is estimated, \( \psi \) is excluded. As a consequence, the DLC-SL-IRT model requires only a one-dimensional integration.

**Specification and Estimation of ILC-IRT Models.** The estimation of ILC-IRT models in the MM-IRT framework requires response times to be organized in a diagonal format (Table 2). On the basis of this data setup, the measurement structure of typical time expenditure is expressed by modifying Equation (7) to

\[
(l_{ij}|C_{ij} = 1) = \nu_k + \lambda_k q_{ij} + \varepsilon_{ijk}, \quad (19)
\]

and Equation (8) to

\[
(l_{ij}|C_{ij} = 0) = \tilde{\nu}_k + \tilde{\lambda}_k q_{ij} + \tilde{\varepsilon}_{ijk}. \quad (20)
\]

The index \( k \) in Equations (19) and (20) replaces the item index \( j \), which means that \( \nu_k = \nu_j, \tilde{\nu}_k = \tilde{\nu}_j, \lambda_k = \lambda_j, \tilde{\lambda}_k = \tilde{\lambda}_j, \sigma_{\varepsilon}^2 = \sigma_{\varepsilon}^2, \) and \( \sigma_{\varepsilon}^2 = \sigma_{\varepsilon}^2 \) for \( j = k \). The term \( q_{ij} \) refers to the individual \( i \)’s scores on a node variable that is fully determined by the \( J \times 1 \) unit vector \( u_i \) with effects that are allowed to vary on the between-individual level, so that

\[
q_{ij} = \xi_i u_{ij}, \quad (21)
\]

where the between-individual random effect variable \( \xi \) is specified to follow a standard normal distribution for model identification \([\xi \sim N(0, 1)]\). The difference between the two ILC-IRT model variants is whether the log response times’ loadings interact (ILC-RE-IRT) or do not interact (ILC-RI-IRT) with the latent class variable. In the case of the ILC-RE-IRT model the constraints \( \tilde{\lambda}_k = 0 \) for all \( k = 1, 2, \ldots, K \) are applied, whereas in the case of the ILC-RI-IRT model, the constraints \( \lambda_k = \tilde{\lambda}_k \) for all \( k = 1, 2, \ldots, K \) are imposed.
The MML estimation of ILC-IRT models is more challenging. The log likelihood to be maximized is given by (e.g., Vermunt, 2008)

\[
\log L = \sum_{i=1}^{N} \log L_i(Y_i, T_i; \phi),
\]

which means that, in these models, we consider the joint distribution of item responses and response times that have a probability density that corresponds to

\[
L_i(Y_i, T_i; \phi) = \int \prod_{j=1}^{J} P(y_{ij}|\xi, \eta; \phi) f(l_{ij}|\xi, \theta, \eta; \phi) g(\xi, \theta, \eta; \phi) d\xi d\eta d\theta,
\]

Illustrative Application to the PIAAC Data

Illustrative Data. PIAAC is an international large-scale assessment of proficiencies in adults. Here, we consider the literacy test administered in the French-speaking samples (France and Canada). We chose this sample because previous research suggests that it exhibits a typical level of disengagement (Goldhammer et al., 2016), and it offers a sample size that is large enough to estimate the models. In PIAAC, a two-stage booklet design is used. In the first stage, examinees are randomly assigned to one of three booklets. Each booklet has a different level of difficulty and consists of nine items. In stage two, individuals are then assigned to one of four different booklets that contain 11 items each. All items have an open response format, which means that the success probability of disengaged responses can be expected to be essentially zero. In order to derive a sample with a decent number of items without introducing additional complications caused by the design, we selected a subgroup of examinees who had worked on a specific combination of stage-one and stage-two booklets (“low” and “intermediate one”). To keep the examples simple, we used only examinees who answered all 20 items (\(N = 637\)).

Model Specification and Estimation. All IRT models for response engagement were estimated with Mplus 8.4 (Muthén & Muthén, 2017) using the accelerated EM algorithm and employing 15 integration points per dimension. As models that include latent class variables could be prone to local optima (Lubke & Muthén, 2005), we employed multiple random starting values to check whether the best log likelihood had been replicated. The Mplus input files for all models are provided in the Supplemental Material.

In our application of DLC-IRT models, the \(\tilde{\beta}\)-parameters were fixed to imply a zero chance probability (i.e., \(\tilde{\beta}_k = 15\) for all \(k = 1, 2, ..., K\)). We included this
constraint because there were no correct responses that were classified as disengaged. In specifying the ILC-IRT models, we constrained the $\beta$-parameters to be equal to each other, and we assumed a homoscedastic residual variance of disengaged log response times. The reason for these constraints was that the data set appeared too small to provide solid item-specific parameter estimates given the rather low prevalence of disengaged responses. In line with the previous arguments according to which equal measurement intercepts of disengaged response times could artificially inflate residual variances, we left all measurement intercepts in both response states unconstrained. The estimated $\beta$-parameters indicated a chance probability close to zero (ILC-RE-IRT: $\hat{\beta} = 6.04$, $SE = 0.86$; ILC-RI-IRT: $\hat{\beta} = 5.60$, $SE = 0.79$).

Model-Data Fit. Decisions between DLC-IRT and ILC-IRT models cannot be based on information criteria, but these indices can be used to examine the merits of accounting for response engagement by comparing the fit of each type of IRT model for response engagement with the fit of a suitable reference model. We chose the 2PL IRT model as a reference for the DLC-IRT models, whereas ILC-IRT models were compared to van der Linden’s (2007) speed-accuracy IRT model (denoted as 2PL + RT). Information criteria were furthermore used to examine the different specifications of each type of IRT model for response engagement. Table 3 summarizes the goodness-of-fit statistics of all IRT models. The DLC-SL-IRT model fitted the data better than the 2PL model, and the fit was further improved by extending the model to the DLC-TL-IRT model. Compared to the 2PL + RT model, the model-data fit was improved by the ILC-IRT models. The ILC-RI-IRT model fitted clearly better than the ILC-RE-IRT model.

Response Time and Response Engagement. In both DLC-IRT models the process discrimination parameters documented a strong effect of response times (DLC-SL-IRT: $\hat{\gamma} = 10.28$, $SE = 1.69$, $p < .001$; DLC-TL-IRT: $\hat{\gamma} = 10.26$, $SE = 2.34$, $p < .001$). The response time thresholds estimated in the two models were very similar and

| Model           | # Par. | LL      | AIC     | BIC     | SBIC    |
|-----------------|--------|---------|---------|---------|---------|
| 2PL             | 40     | -6342.1 | 12764.3 | 12942.5 | 12815.5 |
| DLC-SL-IRT      | 61     | -5995.5 | 12113.0 | 12384.8 | 12191.2 |
| DLC-TL-IRT      | 63     | -5970.0 | 12065.9 | 12346.7 | 12146.7 |
| 2PL IRT + RT    | 101    | -18037.3 | 36276.3 | 36726.7 | 36406.0 |
| ILC-RE-IRT      | 146    | -16428.0 | 33147.9 | 33798.6 | 33335.1 |
| ILC-RI-IRT      | 146    | -16308.2 | 32908.3 | 33559.0 | 33095.5 |

Note. LL = model log likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion; SBIC = sample size adjusted BIC.
The DLC-TL-IRT model indicated individual differences in response time thresholds \((\hat{\tau}^t_j)\) that were correlated with proficiency \((\hat{\rho}_{\theta \xi} = -0.80, SE = 0.10, p < .001)\). Table 4 reports the average posterior engagement probabilities for each item. The patterns of item-specific probabilities were similar for both DLC-IRT models, but in the DLC-TL-IRT model, a smaller proportion of responses was classified as engaged (90.4% vs. 92.5%). This difference was due to the higher response time thresholds of less proficient individuals, which caused a larger proportion of responses to be classified as disengaged.

In the ILC-IRT models, the relationships of response engagement with response times are represented by the separation of engaged and disengaged response time distributions. The upper panel in Figure 3 reports the results of the ILC-RE-IRT model on the basis of the engaged and disengaged response time intercepts and the regions.
Figure 3. Separation of log response times of engaged and disengaged responses provided by the independent latent class IRT model with random effects (ILC-RE-IRT) and the independent latent class IRT model with random intercepts (ILC-RI-IRT). Circles represent intercepts of engaged (filled black) and disengaged (open) log response times. Vertical bars represent the range between the 10th and 90th percentile of the residual distributions around the intercepts. Dotted lines visualize the impact of individual differences in typical time expenditure.
around the intercepts that comprised 80% of the values (±1.282 residual standard deviations). Therefore, in the case of engaged responses, the vertical bars indicate the region in which the response times of individuals with an average value of $\xi$ typically occur. Because, in this model, engaged response times reflect between-individual differences due to $\xi$, the engaged response time bands are accompanied by dotted lines. The dotted lines indicate the lower and upper values of the engaged response time bands expected for individuals who scored at either the 10th or the 90th percentile of the $\xi$-distribution. The response time bands of disengaged responses were wide because their residual variance ($\hat{\sigma}^2_{\xi} = 0.74$, $SE = 0.05$) was larger than the residual variances of engaged response times (ranging from $\hat{\sigma}^2_{\xi1} = 0.10$, $SE = 0.01$ to $\hat{\sigma}^2_{\xi7} = 0.35$, $SE = 0.02$). The response time distributions were clearly separated, even at low levels of $\xi$.

The lower panel in Figure 3 describes the response time distributions given by the ILC-RI-IRT model. Because the separation of engaged and disengaged response times is specified to be the same for all individuals, dotted lines that indicate the impact of the random intercept $\xi$ enclose both mixture components. For engaged responses, the intercepts and residual variances (ranging from $\hat{\sigma}^2_{\xi1} = 0.10$, $SE = 0.01$ to $\hat{\sigma}^2_{\xi7} = 0.32$, $SE = 0.02$) of engaged response times were almost identical to the results of the ILC-RE-IRT model, but the intercepts of disengaged response times were higher. Because the residual variance of disengaged responses was large ($\hat{\sigma}^2_{\xi} = 0.71$, $SE = 0.04$), the response time distributions had a larger overlap. This result means that, compared to the ILC-RE-IRT model, in the ILC-RI-IRT model longer response times were less univocally classified as engaged.

The posterior engagement probabilities given by the ILC-IRT models are reported together with the engagement difficulties in Table 4. Overall, the engagement difficulty parameters and the posterior probabilities were similar in both ILC-IRT models. However, the ILC-RI-IRT model indicated a lower proportion of engaged responses (89.0% vs. 90.8%). This difference was reflected in the larger overlap of response time distributions in the ILC-RI-IRT model (Figure 3), which means that, in this model, relatively long response times were more likely to be classified as disengaged than in the ILC-RE-IRT model.

Both ILC-IRT models provided very similar results with respect to the latent characteristics: They estimated a large variability in engagement propensities (ILC-RE-IRT: $\hat{\sigma}_\eta = 4.48$, $SE = 0.30$; ILC-RI-IRT: $\hat{\sigma}_\eta = 4.36$, $SE = 0.29$), and the correlations between typical time expenditures $\xi$, engagement propensities $\eta$, and proficiencies $\theta$ were very similar (ILC-RE-IRT: $\hat{\rho}_{\xi\eta} = .25$, $SE = 0.06$; $\hat{\rho}_{\eta\theta} = .72$, $SE = 0.04$; $\hat{\rho}_{\theta\xi} = -.01$, $SE = 0.05$; ILC-RI-IRT: $\hat{\rho}_{\xi\eta} = .31$, $SE = 0.05$; $\hat{\rho}_{\eta\theta} = .69$, $SE = 0.04$; $\hat{\rho}_{\theta\xi} = .01$, $SE = 0.05$).

Adjustments of IRT Item Parameters. The estimates of the IRT item parameters are summarized in Figure 4. Compared to the 2PL model, both DLC-IRT models estimated lower item discriminations, but discriminations were slightly lower in the DLC-TL-IRT model. This result suggested that the variability in proficiencies
represented by the 2PL model might be inflated by individual differences in response engagement. Figure 4 includes estimates of the item difficulties (i.e., \( \beta_j = \beta_j / \alpha_j \)). The item difficulties estimated by the DLC-IRT models were lower than those estimated by the 2PL model. The only exception was item 15, for which item difficulties need to be interpreted with caution due to a low item discrimination. Overall, the DLC-TL-IRT model resulted in slightly stronger adjustments of item difficulties than the DLC-SL-IRT model. As shown in the lower panels of Figure 4, which refer to the item parameters of the IRT parts of the ILC-IRT models, the item discriminations

Figure 4. Estimates of item discriminations and difficulties derived from DLC-IRT and ILC-IRT models. Results from a conventional 2PL IRT model and a 2PL IRT model including a congeneric measurement model for time expenditure based on log response times (2PL + RT) as a comparison standard.
estimated by the ILC-IRT models were lower than in the 2PL + RT model and were lowest in the ILC-RI-IRT model. With the exception of item 15, which discriminated weakly between proficiency levels, difficulties were estimated to be lower than in the 2PL + RT model. Difficulties were slightly lower in the ILC-RI-IRT model than in the ILC-RE-IRT model.

Adjustments of Proficiency Estimates. The adjustments of the proficiency estimates provided by the IRT models for response engagement relative to the 2PL model were based on EAP estimates of $\theta$ that were, however, given on different metrics. In order to convert the $\theta$-estimates provided by the DLC-IRT and ILC-IRT models to the metric of the 2PL IRT model, we examined the relationships between the proficiency estimates of individuals who were expected to be engaged throughout the test (i.e., the average posterior latent class probabilities were higher than .99). These estimates correlated almost perfectly ($r > .99$) and were used to derive the coefficients of the linear conversion equations, which allowed us to approximate the proficiency metrics of the DLC-IRT and ILC-IRT models to the metric of the 2PL IRT model.

Figure 5 presents the scatter plots of the EAP $\theta$-estimates derived from the 2PL IRT model (x-axis) and converted EAP estimates given by the DLC-IRT models (y-axis). For individuals who were expected to be partially disengaged, both DLC-IRT models provided higher proficiency estimates than the 2PL IRT model. Moreover, the pattern of relationships turned out to be highly similar between the DLC-IRT models, so that, relative to the 2PL IRT model, the $\theta$-estimates were, on average, $d = 0.14$ and $d = 0.15$ standard deviations higher in the DLC-SL-IRT and the DLC-TL-IRT model, respectively. However, in the case of individuals who were expected to be disengaged on essentially all items of the test, the adjustments of the $\theta$-estimates were extreme because their EAP estimates were extremely shrunken to the mean of the proficiency distribution. Apart from these few extreme cases ($N = 10$), the adjustments of the proficiency estimates were reasonable.

Both ILC-IRT models provided proficiency estimates adjusted for response engagement that differed from the estimates of the 2PL IRT model (Figure 5). The marginal distribution of EAP $\theta$-estimates given by the ILC-RE-IRT model was only $d = 0.08$ standard deviations higher than that given in the 2PL IRT model. The adjustment was stronger in the case of the ILC-RI-IRT model ($d = 0.12$). Interestingly, the adjustment of $\theta$-estimates relative to the estimates of the 2PL IRT model in very poorly engaged individuals did not appear to be overly strong, as was the case in the DLC-IRT models. This is because the $\eta$-variable, which exhibited a large variability and was strongly correlated with $\theta$ in both ILC-IRT models, added extra information that prevented the $\theta$-estimates from being overly shrunken to the mean of their distribution.

Summary. All IRT models for response engagement supported the suspicion that the PIAAC literacy assessment was affected by disengaged responses (Goldhammer et al., 2016). In addition, the results indicate that the details of the models’
specifications matter. The preferred DLC-TL-IRT and ILC-RI-IRT models provided stronger adjustments of IRT item parameters and proficiency estimates than the DLC-SL-IRT and ILC-RE-IRT models. However, the information criteria do not tell whether the DLC-TL-IRT or the ILC-RI-IRT model should be chosen as the final model. Nevertheless, we can conclude that DLC-IRT and ILC-IRT models fit the data better than the reference models that do not separate response states and that, within each type of IRT model for response engagement, one parameterization fits the data better than the other. We do not consider this aspect of model selection to

Figure 5. Scatter plots of EAP proficiency estimates provided by the 2PL IRT model (x-axis) and by equated EAP proficiency estimates provided by the DLC-IRT (y-axis in upper panels) and the ILC-IRT models (y-axis in lower panels) for individuals with different levels of engagement (average posterior probabilities of engagement).
be problematic because, in real applications, researchers should decide on one type of IRT model for response engagement on the basis of the perspective they have on response times.

Interestingly, in the present applications, the results pertaining to the item responses and the proficiencies provided by these two preferred DLC-IRT and ILC-IRT models were remarkably similar. Given the fundamental differences between the two types of models, such results should not be taken for granted. Indeed, a closer look at the relationships between response times and posterior engagement probabilities revealed non-monotonic relationships in the ILC-RI-IRT model (i.e., some responses associated with very long response times were classified as disengaged; details available upon request). However, in the present application, only a small number of responses was affected by the non-monotonic relationships, which means that the model estimates were not overly affected by this phenomenon.

Summary and Discussion

IRT models for response engagement are tools that allow item responses and response times to be considered simultaneously, while avoiding the necessarily error-prone procedures for classifying the engagement status of item responses. IRT models for response engagement build upon different assumptions, which means that different IRT models can give different results when they are applied to the same data. Therefore, the objectives of the present article were: (1) to outline the rationale of the commonly employed methods for setting response time thresholds, (2) to describe the assumptions they share with different types of IRT models for response engagement, (3) to suggest model extensions that increase the IRT models’ flexibility, and (4) to provide a flexible MM-IRT framework that makes it possible to estimate various IRT models for response engagement.

Types of IRT Models for Response Engagement

We have delineated that IRT models for response engagement and observed variable procedures are built on common ideas, and we have proposed the distinction between ILC-IRT models, which consider response times as indicators of response engagement, and DLC-IRT models, in which response times serve as predictors of engagement.

Up until today, most IRT models for response engagement belong to the class of ILC-IRT models in which response engagement is defined jointly on the basis of item responses and response times. However, because the continuous log response times carry more information (Wang & Xu, 2015), they are likely to determine the classification of response engagement more strongly than item responses. As such, ILC-IRT models are connected to the VIB method, which uses only response times, but this connection is loose because ILC-IRT models use the complete data and are based on additional structural assumptions. In our view, ILC-IRT models can be best
understood as extensions of van der Linden’s (2007) speed-accuracy IRT model, which make it possible to account for violations of the constant time expenditure assumption due to disengaged responses. Disengaged item responses are specified to not depend on proficiency and to be accompanied by response times that stem from a different distribution than that of engaged responses. In ILC-IRT models response times contribute to the classification of response engagement according to their typicality for each engagement state (Equation 12), which means that these models could classify very slow responses as disengaged responses. From the perspective of the typical time expenditure assumption, such kinds of non-monotonic relationships make sense because both fast and very slow responses could be considered to be atypical and disengaged.

The variants of ILC-IRT models presented here use response times differently for classifying response engagement. ILC-RE-IRT models assume that engaged response times differ from an unstructured reference distribution, which means that the more strongly an item’s response time deviates from the reference distribution, the more likely it is that the response is engaged. In contrast, ILC-RI-IRT models state that both engaged and disengaged responses reflect individuals’ typical time expenditure and that the separation between engaged and disengaged response times is the same for all individuals. As shown in the exemplary application, the variants of ILC-IRT models can provide different results. Therefore, researchers need to carefully consider which type of ILC-IRT model is more plausible. The ILC-RI-IRT model might be a better choice when tests consist of items that require nontrivial time expenditure even when responses are not engaged.

By relating response times to response engagement based on the theoretical consideration that faster responses can—with higher certainty—be considered to be disengaged, and without recurring on assumptions about the structure of response times, DLC-IRT models closely mirror the perspectives taken in the research on rapid guessing behavior. Therefore, although DLC-IRT models have not attracted much attention in the literature (but see Pokropek, 2016), we believe that they have much to offer. In this article, we have proposed two extensions of Pokropek’s (2016) DLC-IRT model that both include item-specific response time thresholds that are based on a rationale similar to that of the combined VICS and VII methods (Wise, 2019). As such, DLC-IRT models can be conceived as model-based, one-step procedure counterparts of the VICS-VII procedure.

Of the two variants of DLC-IRT models presented, the DLC-SL-IRT model is closest to the typical approach of classifying each individual’s response engagement on the basis of the same response time thresholds. The DLC-TL-IRT is an extension thereof that allows for individual differences in response time thresholds that could be correlated with proficiency. The failure to include such a component in DLC-IRT models has been discussed as a potential source of bias (Pokropek, 2016; Rios et al., 2017), and we have demonstrated that including individual differences in response time thresholds could indeed affect item parameter and proficiency estimates.
Taken together, the DLC-IRT and ILC-IRT models are conceptually different and fit to different theoretical perspectives of response times as reflections or causes of response engagement. Therefore, decisions in favor of one type of model should optimally be theoretically justified, or the substantive considerations underlying the decision should at least be clearly laid out. Such explanations should be self-evident in research. In the present context, they gain additional importance because DLC-IRT and ILC-IRT models use the input data in different ways so that decisions cannot be guided by information criteria.

The MM-IRT Framework

A further objective of this article was to describe an MM-IRT framework that allows a variety of IRT models for response engagement to be estimated by means of the Mplus software. Our framework is based on a diagonal arrangement of the latent variables’ indicators. This setup combines the strengths of traditional wide and long data formats. An advantage of the wide data format is that all measurement parameters can differ between indicators and latent classes. However, this flexibility comes at the price of requiring a large number of latent class variables, which limits the application of IRT models for response engagement to rather small numbers of items (e.g., Molenaar et al., 2016). In the long data format, multiple latent class variables can be replaced with one class variable located on the within-individual level. The measurement parameters of the indicators of the latent continuous variables are modeled by the effects of the item indicators and their interactions with the latent class variable (Pokropek, 2016). The drawbacks of this arrangement are that it is less flexible in modeling residual structures and that the estimation of item-specific discrimination and loading parameters is computationally demanding. The diagonal data arrangement provides the same flexibility as the wide format and allows a large number of items to be included in the models, while still supporting computationally efficient MML estimation.

The MM-IRT framework provides great flexibility because, in principle, all parameters can be specified to be item-specific. The IRT models applied in this article include fewer constraints than the models employed in the literature (e.g., Ulitzsch et al., 2020), and the constraints can be further relaxed. An important aspect of the MM-IRT framework is that it is not limited to item response times, and it allows alternative predictors or indicators of response engagement to be incorporated into IRT models for response engagement. This feature makes it possible to investigate the utility of other within-individual variables, such as interest ratings (e.g., Lindner et al., 2019) or repeating response patterns (e.g., Cui, 2020), as alternative or additional predictors or indicators of response engagement in DLC-IRT and ILC-IRT models, respectively. Similarly, the proposed models can be extended by between-individual variables, which makes it possible to study their relationships with individual differences in thresholds or engagement (e.g., Nagy et al., 2019). These possibilities should be explored in more detail in further research.
Further Research

The models presented here are, in some respects, more constrained than some other models suggested in the literature. The IRT model applied to engaged responses does not include pseudo-guessing parameters, which might be desired in the case of multiple choice items. We plan to explore this modeling option in future studies.

A concern in the context of ILC-IRT models could be the assumption that item-specific engagement states are conditionally independent. An alternative is to consider that item-specific engagement states depend on previous states. Molenaar et al. (2016) suggested a model in which response processes are modeled by a hidden Markov process, but their model has not been applied to response engagement yet. Future research could investigate possibilities to include such dependencies in IRT models for response engagement. However, an interesting line of research would be to study whether serial dependencies can be accounted for by means of DLC-IRT models that are computationally less demanding. In this vein, more research could be devoted to DLC-IRT models to increase their flexibility, for example, by including item-specific effects of response times and accounting for measurement error in response times (Molenaar & de Boeck, 2018).

Conclusion

The MM-IRT framework makes it possible to specify and estimate a variety of ILC-IRT and DLC-IRT models that extend the arsenal of latent variable models for response engagement. Furthermore, the MM-IRT framework makes the IRT models for response engagement accessible to applied researchers. Indeed, as shown in the Supplemental Material., the Mplus input code for specifying the models is not complicated. We hope that the ideas and the procedures outlined here will prove useful to researchers studying the impact of response engagement (Wise, 2015) and will stimulate research that extends our knowledge about test-taking motivation (e.g., Nagy et al., 2018).

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 iDs

Gabriel Nagy https://orcid.org/0000-0001-6897-212X
Esther Ulitzsch https://orcid.org/0000-0002-9267-8542
**Supplemental Material**

Supplemental material for this article is available online.

**Notes**

1. Note that Ulitzsch et al.’s (2020) ILC-IRT model includes missing data indicators in addition to item responses and response times. However, when missing data indicators are excluded, their model is included in the class of ILC-IRT models presented in this article.

2. The PDF values associated with a response time $l_{ij}$ are a function of the individual $i$’s expected response times to item $j$ in the engaged and disengaged state, as well as the variability around the expectations. Expected response times correspond to $v_j + \lambda_j \xi_i$ and $\bar{v}_j + \lambda_j \xi_i$ (Equations 7 and 8) for the engaged and disengaged state, respectively. The variability around the expectations is represented by the residual variances $\sigma^2_e$ and $\sigma^2_{\xi}$. 

3. Note that DLC-IRT models do not per se require the $\beta$-parameters to be fixed. We investigated this issue with a small simulation study in which the chance level of success equaled $P(v_{ij} = 1 | C_{ij} = 0) = 0.10$ for all items, and we found that the parameter was essentially unbiased.

**References**

Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 27–51). Information Age Publishing.

Cui, Z. (2020). On a new algorithm for removing repeating patterns in similarity analysis. *Educational and Psychological Measurement*, 80, 446–460. https://doi.org/10.1177/0013164419882980

Darrell Bock, R., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179–197. https://doi.org/10.1007/bf02291262

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, 10, 102. https://doi.org/10.3389/fpsyg.2019.00102

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD education working papers no. No. 133). OECD Publishing.

Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing individual differences in basic computer skills. *European Journal of Psychological Assessment*, 29, 263–275. https://doi.org/10.1027/1015-5759/a000153

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Pack, I. (2016). A new procedure for detection of students’ rapid guessing responses using response time. *Applied Measurement in Education*, 29, 173–183. https://doi.org/10.1080/08957347.2016.1171766

Klein Entink, R. H., Kuhn, J. T., Hornke, L. F., & Fox, J. P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, 14, 54–75. https://doi.org/10.1037/a0014877

Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students’ test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, 2, 1–24. https://doi.org/10.1186/s40536-014-0008-1

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology*, 10, 1533. https://doi.org/10.3389/fpsyg.2019.01533
List, M. K., Robitzsch, A., Lüdtke, O., Köller, O., & Nagy, G. (2017). Performance decline in low-stakes educational assessments: Different mixture modeling approaches. Large-scale Assessments in Education, 5, 5. https://doi.org/10.1186/s40536-017-0049-3

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. Psychological Methods, 10(1), 21–39. https://doi.org/10.1037/1082-989x.10.1.21

Lu, J., Wang, C., Zhang, J., & Tao, J. (2020). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. British Journal of Mathematical and Statistical Psychology, 73, 261–288. https://doi.org/10.1111/bmsp.12175

Molenaar, D., & de Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. Psychometrika, 83, 279–297. https://doi.org/10.1007/s11336-017-9602-9

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. Multivariate Behavioral Research, 51, 606–626. https://doi.org/10.1080/00273171.2016.1192983

Muthén, L. K., & Muthén, B. O. (2017). Mplus user’s guide (8th ed.). Muthén & Muthén.

Nagy, G., Nagengast, B., Becker, M., Rose, N., & Frey, A. (2018). Item position effects in a reading comprehension test: An IRT study of individual differences and individual correlates. Psychological Test and Assessment Modeling, 60, 165–187.

Nagy, G., Nagengast, B., Frey, A., Becker, M., & Rose, N. (2019). A multilevel study of position effects in PISA achievement tests: Student- and school-level predictors in the German tracked school system. Assessment in Education Principles Policy and Practice, 26(4), 422–443. https://doi.org/10.1080/0969594x.2018.1449100

OECD. (2017). PISA 2015 technical report. https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. Journal of Educational and Behavioral Statistics, 41(3), 300–325. https://doi.org/10.3102/1076998616636618

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? International Journal of Testing, 17, 74–104. https://doi.org/10.1080/15305058.2016.1231193

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. Journal of Educational Measurement, 34(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Ulitzsch, E., Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. British Journal of Mathematical and Statistical Psychology, 73, 83–112. https://doi.org/10.1111/bmsp.12188

van Barneveld, C., Pharand, S.-L., Ruberto, L., & Haggarty, D. (2013). Student motivation in large-scale assessments. In M. Simon & K. Ercikan, Michel Rousseau (Eds.), Improving large-scale assessment in education: Theory, issues and practice (pp. 43–61). Routledge.

Van Den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. Journal of Educational and Behavioral Statistics, 28, 369–386. https://doi.org/10.3102/10769986028004369

van der Linden, W. J. (2006). A lognormal model for response times on test items. Journal of Educational and Behavioral Statistics, 31, 181–204. https://doi.org/10.3102/10769986031002181
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on
test items. *Psychometrika*, 72, 287–308. https://doi.org/10.1007/s11336-006-1478-z

Vermunt, J. K. (2003). 7. Multilevel latent class models. *Sociological Methodology*, 33, 213–239. https://doi.org/10.1111/j.0081-1750.2003.t01-1-00131.x

Vermunt, J. K. (2008). Multilevel latent variable modeling: An application in education
testing. *Austrian Journal of Statistics*, 37, 285–299. https://doi.org/10.17713/ajs.v37i3&4.309

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response
accuracy. *British Journal of Mathematical and Statistical Psychology*, 68, 456–477. https://
doi.org/10.1111/bmsp.12054

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item
preknowledge: A comparison of mixture modeling method and residual method. *Journal of
Educational and Behavioral Statistics*, 43(4), 469–501. https://doi.org/10.3102/107699861876123

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data.
*Applied Measurement in Education*, 28, 237–252. https://doi.org/10.1080/08957347 .2015.1042155

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and
implications. *Educational Measurement Issues and Practice*, 36, 52–61. https://doi .org/10.1111/emip.12165

Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds.
*Applied Measurement in Education*, 32, 325–336. https://doi.org/10.1080/08957347 .2019.1660350

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-
moderated IRT model. *Journal of Educational Measurement*, 43, 19–38. https://doi .org/10.1111/j.1745-3984.2006.00002.x

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation
in computer-based tests. *Applied Measurement in Education*, 18, 163–183. https://doi .org/10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (Eds.). (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S. L., & Smith, L. F. (2016). The validity of assessment when students don’t give good
effort. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 204–220). Routledge.