the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2020), 73, 83–112

www.wileyonlinelibrary.com

Editor's Choice

A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response

Esther Ulitzsch[1]*, Matthias von Davier[2] and Steffi Pohl[1]

[1]Methods and Evaluation/Quality Assurance, Freie Universität Berlin, Germany
[2]National Board of Medical Examiners, Philadelphia, Pennsylvania, USA

In low-stakes assessments, test performance has few or no consequences for examinees themselves, so that examinees may not be fully engaged when answering the items. Instead of engaging in solution behaviour, disengaged examinees might randomly guess or generate no response at all. When ignored, examinee disengagement poses a severe threat to the validity of results obtained from low-stakes assessments. Statistical modelling approaches in educational measurement have been proposed that account for non-response or for guessing, but do not consider both types of disengaged behaviour simultaneously. We bring together research on modelling examinee engagement and research on missing values and present a hierarchical latent response model for identifying and modelling the processes associated with examinee disengagement jointly with the processes associated with engaged responses. To that end, we employ a mixture model that identifies disengagement at the item-by-examinee level by assuming different data-generating processes underlying item responses and omissions, respectively, as well as response times associated with engaged and disengaged behaviour. By modelling examinee engagement with a latent response framework, the model allows assessing how examinee engagement relates to ability and speed as well as to identify items that are likely to evoke disengaged test-taking behaviour. An illustration of the model by means of an application to real data is presented.

1. Introduction

The aim of large-scale assessments (LSAs) is to measure examinee competencies using test items. In doing so, it is assumed that examinees actively try to determine the correct answer to every item by employing their abilities (Schnipke & Scrams, 1997; Wang & Xu, 2015). Most comparative LSAs, however, are low-stakes for examinees and aim at system-level comparisons. As such, examinee test performance in most LSAs has few or no consequences for examinees themselves and examinees may not be fully engaged when attempting the items. When disengaged, examinees might attempt items without

*Correspondence should be sent to Esther Ulitzsch, Methods and Evaluation/Quality Assurance, Freie Universität Berlin, Habelschwerdter Allee 45, Berlin 14195, Germany (email: esther.ulitzsch@fu-berlin.de).

DOI:10.1111/bmsp.12188

applying their abilities, but instead proceed quickly through the assessment by randomly guessing on multiple-choice (MC) items, answering items with an open-response (OR) format only perfunctorily, or generating no response at all (Verbić & Tomić, 2009; Wise & Gao, 2017). Such disengaged test-taking behaviour poses a severe threat to the validity of results obtained from LSAs since test scores assumed to reflect the level of competency may be confounded with the level of disengagement (Braun, Kirsch, & Yamamoto, 2011). Identifying and understanding the processes associated with examinee disengagement is therefore paramount for drawing valid inferences on examinee ability.

In this study, we argue that both rapid guesses and item omissions can be understood as indicators of examinee disengagement (see Wise & Gao, 2017). To capture the underlying processes, we bring together research on modelling examinee engagement and research on item-level non-response and provide a generalized modelling framework that identifies disengagement by jointly considering information on responses, omissions, and response times (RTs).

The remainder of this article is structured as follows: First, we review current approaches for identifying examinee disengagement as well as for handling item omissions. Second, we present a hierarchical latent response framework for examinee disengagement in terms of guessing and omitting. We then evaluate the statistical performance of the proposed model, illustrate how it differs from current approaches for identifying examinee disengagement and handling item omissions, and illustrate its application employing data from the Programme for International Student Assessment (PISA) 2015.

2. Previous approaches for identifying and handling disengaged behaviour

2.1. Guessing and perfunctory answers

Previous approaches have conceptualized examinee disengagement in terms of rapid guesses on MC items and perfunctory answers on OR items. Such disengaged observed item responses typically show measurement properties that differ from those of engaged responses (Cao & Stokes, 2008; Meyer, 2010; Schnipke & Scrams, 1997; Wang & Xu, 2015; Yamamoto & Everson, 1997). As a result, not considering that a portion of observed responses may stem from disengaged test-taking behaviour potentially yields biased and less efficient person and item parameter estimates (Cao & Stokes, 2008; Pokropek, 2016; Rios, Guo, Mao, & Liu, 2017; Wang & Xu, 2015). To mitigate these challenges, various procedures for identifying and filtering disengaged responses have been suggested (Bhola, 1994; Goldhammer, Martens, Christoph, & Lüdtke, 2016; Schnipke, 1996; Schnipke & Scrams, 1997; Wang & Xu, 2015; Wise & DeMars, 2005, 2006).

2.1.1. Response-time-based scoring techniques

In RT-based scoring methods for identifying and filtering disengaged responses, responses associated with RTs below a certain threshold are considered to be rapid guesses. Different approaches exist for establishing these thresholds. The most heuristic threshold method is to define a common threshold for all items representing the minimum amount of time needed to give an engaged response (Wise, Kingsbury, Thomason, & Kong, 2004). Item-specific thresholds can be established by setting the threshold to, for example, 10% of the average time (Wise & Ma, 2012), by visually assessing bimodal RT distributions for a distinctive gap (Wise, Pastor, & Kong, 2009), or by assessing RT distributions jointly with

the conditional proportion correct in order to identify an RT threshold at which accuracy exceeds what would be expected from random responding (Goldhammer *et al.*, 2016; Guo *et al.*, 2016; Lee & Jia, 2014).

2.1.2. Model-based approaches

Model-based approaches aiming to identify disengaged test-taking behaviour usually apply mixture modelling techniques, with responses, and, if considered, RTs assumed to stem from two different processes: solution behaviour and rapid guessing behaviour. For responses stemming from solution behaviour, customary item response theory (IRT) models are assumed. That is, probability correct is modelled as a function of examinee ability and item difficulty. Responses stemming from rapid guessing processes are assumed to contain no information on ability; probability correct under disengaged behaviour is thus assumed to correspond to the probability of guessing correctly at chance level (Schnipke & Scrams, 1997; Wang & Xu, 2015). RTs are either assumed to stem from different lognormal distributions with different means and variances associated with solution and random guessing behaviour (Meyer, 2010; Schnipke & Scrams, 1997; Wang & Xu, 2015) or employed to predict latent class membership (Pokropek, 2016).

2.1.3. Assumptions and limitations

RT-based scoring techniques for identifying examinee disengagement are rather heuristic and might considerably disagree in the rate of responses classified as rapid guesses (Lee & Jia, 2014) or perfunctory answers (Goldhammer *et al.*, 2016). For instance, Goldhammer *et al.* (2016) have reported proportions of perfunctory answers ranging from 0.05% to 8.20% for different threshold methods applied to the same data set. Mixture models for disengaged behaviour, on the other hand, often come with strong assumptions regarding the processes underlying examinee disengagement. In mixture models for disengaged behaviour, mixing proportions are allowed at the population (Meyer, 2010), examinee (Cao & Stokes, 2008; Mislevy & Verhelst, 1990; Wang & Xu, 2015), item (Schnipke & Scrams, 1997) or item-by-examinee level (Pokropek, 2016). While models allowing for varying mixing proportions at the item level assume that items can evoke disengaged behaviour to a different degree, they assume all examinees to be equally prone to show disengaged behaviour. Models assuming examinee-specific mixing proportions allow for the probability of being disengaged to vary across examinees, while the proportion of disengaged responses is assumed to be constant across items. The probability of disengaged responses, however, has repeatedly been shown to be related to both examinee characteristics such as academic ability or achievement goals and item characteristics such as response format or position (Goldhammer *et al.*, 2016; Lee & Jia, 2014; Wise *et al.*, 2009). Considering this when modelling examinee engagement renders it necessary to allow for mixing proportions at the item-by-examinee level. To our knowledge, the grade of membership framework presented by Erosheva (2002) and adapted for identifying examinee disengagement by Pokropek (2016) is the only framework that allows for mixing proportions at the item-by-examinee level. It does so by regressing item-by-examinee-level mixing proportions on the associated RTs.

In addition, mixture models for identifying examinee disengagement do not model the probability of being engaged jointly with ability but rather as an independent process. Thus, these models assume ability and engagement to be unrelated constructs. In RT-based scoring approaches, on the other hand, item responses identified to be the result of

guessing behaviour are often coded as missing and therefore ignored when estimating ability. Doing so comes with the assumption that the missing responses induced through such filtering techniques are ignorable in the sense that they are missing at random (MAR) given the observed (engaged) responses and the background variables considered, and that the processes leading to disengaged item responses are unrelated to ability (Pokropek, 2016; Rios *et al.*, 2017; Rubin, 1976). A rich body of research, however, suggests that motivation and the tendency to show guessing behaviour are indeed related to ability (Boe, May, & Boruch, 2002; Braun *et al.*, 2011; Goldhammer *et al.*, 2016; Wise & DeMars, 2005; Wise *et al.*, 2009). Not taking this into account has been shown to yield biased ability estimates (Pokropek, 2016; Rios *et al.*, 2017). To overcome these limitations, there is a need for a model-based approach that allows for the probability of observing disengaged behaviour to vary at the item-by-examinee level, as well as joint modelling of the processes underlying examinee disengagement and ability.

2.2. *Omissions*

Various studies have related the occurrence of item omissions to lack of examinee motivation (Cosgrove, 2011; Jakwerth & Stancavage, 2003; Köhler, Pohl, & Carstensen, 2015a; Verbić & Tomić, 2009; Wise & Gao, 2017). Decline in test scores over time, for instance, has been attributed to a decline in examinee motivation, with an increase in omission rates taken as an indicator of examinee disengagement (Cosgrove, 2011; Sachse, Mahler, & Pohl, 2019). Likewise, it has been suggested to employ the rate of item omissions on background questionnaires as an indicator of disengagement in cognitive assessments, with the rationale being that examinees who are not motivated to fill out the background questionnaire might also be less motivated to engage with the items of the cognitive assessment (Boe *et al.*, 2002).

Notwithstanding, there is an ongoing discussion about the treatment of item omissions in the cognitive assessments of LSAs. Operationally in LSAs there is considerable variety in the treatment of item omissions, where omissions are either ignored, scored as incorrect, or scored as partially correct (see Pohl, Gräfe, & Rose, 2014, for an overview). While scoring item omissions as wrong assumes the probability of a correct response to an omitted item to be zero (Rose, von Davier, & Xu, 2010), ignoring item omissions implies ignorability (Rose *et al.*, 2010). In the case that ignorability does not hold, ignoring missing data jeopardizes validity of inference and can induce bias to person and item parameter estimates (de Ayala, Plake, & Impara, 2001; Culbertson, 2011; Finch, 2008; Köhler, Pohl, & Carstensen, 2015b, 2017; Pohl *et al.*, 2014; Rose, 2013; Rose *et al.*, 2010).

2.2.1. Response-time-based scoring techniques

RT-based scoring techniques for item omissions aim to distinguish item omissions occurring due to processes different from and similar to those operating when examinees generate (engaged) responses. For item omissions associated with RTs remarkably shorter than RTs associated with observed responses, it is assumed that the examinee did not engage with the item but skipped it without trying to solve it. Item omissions associated with RTs that do not notably differ from RTs associated with (wrong) observed responses are assumed to have occurred for skill-related reasons, since the examinee engaged sufficiently long with the item to generate a response, but decided not to. To distinguish between these two types of omissions, the Programme for the International Assessment of

Adult Competencies (PIAAC) employs a 5-s scoring rule, where item omissions associated with RTs exceeding 5 s are treated as wrong. Otherwise, item omissions are considered not attempted and treated as missing responses in all further analyses (Yamamoto, Khorramdel, & von Davier, 2013). Recent approaches for RT-based scoring of omitted responses extend this rationale by allowing for item-specific, empirically derived thresholds (Frey, Spoden, Goldhammer, & Wenzel, 2018; Weeks, von Davier, & Yamamoto, 2016).

2.2.2. Model-based approaches

In model-based approaches for non-ignorable item omissions, the missingness mechanism assumed to underlie item omissions is usually modelled via an additional manifest or latent variable which represents the examinees' propensity to omit items (Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999; Rose *et al.*, 2010), such that response and omission behaviour are modelled jointly.

For response indicators u_{ij}, representing person i's response on item j, customary IRT models are employed, with the probability of a correct response being modelled as a function of ability θ_i and item difficulty b_j:

$$p(u_{ij} = 1) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)}. \tag{1}$$

Omission indicators d_{ij} contain information on whether examinee i generated a response to item j, with 1 indicating an item omission and 0 an observed response. The probability of an item omission is modelled as

$$p(d_{ij} = 1) = \frac{\exp(\xi_i - a_j)}{1 + \exp(\xi_i - a_j)}, \tag{2}$$

with ξ_i denoting person omission propensity and a_j item omission difficulty. A multivariate normal distribution is assumed for ability and omission propensity.

Traditionally, model-based approaches have relied only on information as to whether a response has been observed or not. Based on the work of Pohl, Ulitzsch, and von Davier (2019), Ulitzsch, von Davier, and Pohl (2019) have extended model-based approaches for non-ignorable item omissions by integrating them with models for RTs, allowing for different processes determining the time examinees require to generate a response or to omit an item. Doing so allows assessment of the degree to which these processes differ and, as such, for a finer-grained understanding of the occurrence of item omissions as well as test-taking behaviour in general.

2.2.3. Assumptions and limitations

Although RT-based scoring techniques for item omissions allow different types of item omissions to be distinguished, they assume that either item omissions are ignorable or the probability of solving an omitted item is zero (Lord, 1983; Rose, 2013). By modelling omission propensity jointly with ability, model-based approaches for item omissions overcome these assumptions, allow to assess how examinee ability relates to the probability of omitting responses, and have been shown to yield unbiased item and person parameter estimates, even when the missingness mechanism is non-ignorable in the sense

that parameters of the response model are not distinct from those of the missingness model (Holman & Glas, 2005; Pohl *et al.*, 2014; Rose *et al.*, 2010; Ulitzsch *et al.*, 2019). If one were to consider omissions as indicators of disengaged behaviour while assuming that all observed responses stem from solution behaviour and that examinees do not omit while engaged, the omission propensity in these models can be understood as an examinee disengagement parameter that is modelled jointly with ability. As such, these models overcome the assumption of independence between the processes governing disengaged behaviour and ability inherent to model-based approaches for disengaged guessing. They are, however, restrictive in that they assume all item omissions to stem from the same data-generating processes and all observed responses to stem from engaged response processes.

3. Proposed model

Conceptualizing disengaged test-taking behaviour in terms of both randomly guessing (or producing perfunctory answers) and omitting, we present a hierarchical latent response model for identifying and modelling the processes associated with examinee disengagement jointly with the processes associated with engaged responses.[1] We thereby bring together research on examinee disengagement and non-response behaviour. Addressing limitations of previously developed approaches, the speed-accuracy + engagement (SA+E) model allows for item-by-examinee-specific engagement probabilities, defines engagement in terms of both random guessing (or perfunctory answers) and disengaged item omissions, and models processes associated with examinee disengagement jointly with ability. To that end, we employ mixture models that identify disengagement at the item-by-examinee level by assuming different data-generating processes underlying item responses, omissions, and RTs associated with engaged and disengaged behaviour. Item-by-examinee mixing proportions are modelled with a latent response framework employing an IRT model. The framework is shown in Figure 1, where the left- and right-hand parts depict the models for disengaged and engaged behaviour, respectively.

Following Wang and Xu (2015), latent engagement indicators Δ_{ij} denote whether examinee i has engaged in solution behaviour when attempting item j or not, with 0 and 1 indicating disengaged and solution behaviour, respectively. Whether or not examinee i generated an engaged response to item j is not observable. Engaged and disengaged behaviours, however, are assumed to result in different distributions of item responses, omissions, and RTs.

3.1. Engaged behaviour

When attempting items in an engaged manner, examinees are assumed to generate engaged responses to all items attempted. That is, if $\Delta_{ij} = 1$ the probability of a correct response on response indicator u_{ij} is assumed to be a function of person ability θ_i and the item's difficulty b_j. In line with the frameworks of analysis implemented in major LSAs such as PISA (OECD, 2017), we present the framework employing a Rasch model for item responses as given by equation (1).

[1] Note that in the proposed model, examinee disengagement is defined in terms of both guessing (or perfunctory answers) and omitting. The model can easily be simplified to assuming that disengaged examinees either only guess or only omit in the case where no omissions are observed, or for theory-based reasons all observed responses can be assumed to stem from engaged response processes. In the Supporting Information, simplified versions of the model are presented and their relationship to other state-of-the-art approaches for item omissions and disengaged guesses are discussed.

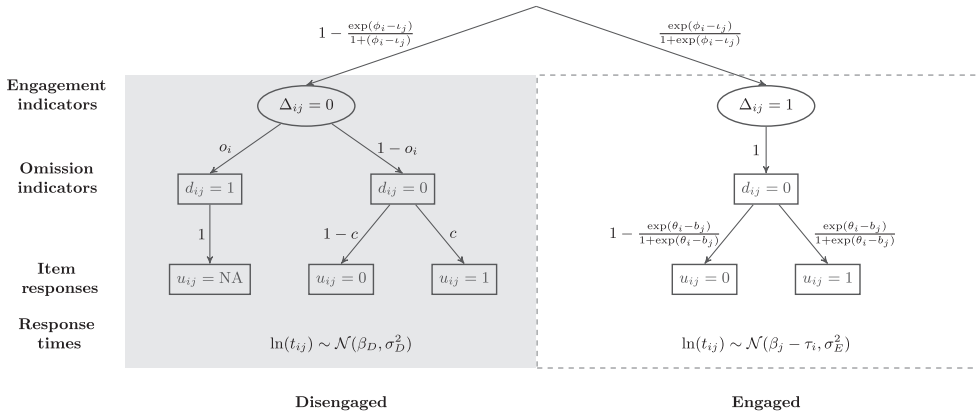

Figure 1. Hierarchical latent response SA+E framework.

Following van der Linden (2007) and Wang and Xu (2015), RTs t_{ij}, denoting the time examinee i interacted with item j, are assumed to follow a lognormal distribution governed by examinee working speed τ_i and item time intensity β_j when associated with an engaged response:

$$\ln(t_{ij}|\Delta_{ij}=1)\sim\mathcal{N}(\beta_j-\tau_i,\sigma_{\mathrm{E}}^2). \tag{3}$$

For reasons of simplicity, we assume a common residual variance σ_{E}^2 (van der Linden, 2007).

Item omissions are assumed not to occur when examinees are engaged. We therefore fix the probability of observing an item omission to zero if $\Delta_{ij}=1$. Thus,

$$p(d_{ij}=1|\Delta_{ij}=1)=0. \tag{4}$$

Conversely, this restriction corresponds to the assumption that examinee disengagement is observable in the case an item is omitted.

3.2. Disengaged behaviour

When disengaged ($\Delta_{ij}=0$), we assume that examinees either randomly guess or omit. Whether examinees omit or guess is modelled via an examinee-specific but not item-specific omission probability o_i which describes the probability that examinee i omits ($d_{ij}=1$) rather than guesses ($d_{ij}=0$) when attempting an item in a disengaged manner. o_i is modelled as a function of ability θ_i and speed τ_i via a logistic regression, thereby allowing for differences in omission behaviour depending on the examinee's ability and speed level:

$$p(d_{ij}=1|\Delta_{ij}=0)=o_i=\frac{\exp(\gamma_0+\gamma_1\theta_i+\gamma_2\tau_i)}{1+\exp(\gamma_0+\gamma_1\theta_i+\gamma_2\tau_i)}. \tag{5}$$

For observed disengaged responses, the probability of a correct guess is assumed to be determined by a common guessing parameter c (Schnipke & Scrams, 1997; Wang & Xu, 2015):

$$p(u_{ij} = 1|\Delta_{ij} = 0) = c. \tag{6}$$

Following Schnipke and Scrams (1997), we assume that neither person nor item characteristics affect the distribution of RTs when examinees are disengaged and produce responses by guessing or omit items. Thus, under $\Delta_{ij} = 0$, RTs for all items and examinees are assumed to follow a lognormal distribution governed by a common mean across all items and examinees β_D and variance σ_D^2:

$$\ln(t_{ij}|\Delta_{ij} = 0) \sim \mathcal{N}(\beta_{\mathrm{D}}, \sigma_{\mathrm{D}}^2). \tag{7}$$

In the proposed framework, it is assumed that examinees tend to require less time to interact with an item when disengaged than to read, understand, and generate an engaged response to the item (Wise, 2017). We incorporate this assumption by assuming that all time intensities for the RTs associated with engaged behaviour β_j are the sum of the common mean β_{D} and an item-specific, positive offset parameter β_j^*. That is,

$$\beta_j = \beta_{\mathrm{D}} + \beta_j^*, \quad \text{where } \beta_j^* \geq 0. \tag{8}$$

The offset parameter β_j^* indicates how much longer examinees need to engage with the item to generate an engaged response rather than to omit or guess.

3.3. Higher-order models

Whether examinee i engaged in solution behaviour when attempting item j is only partially observable; however, it determines the measurement properties of the observed responses and associated RTs. Engagement indicators Δ_{ij} thus represent latent response variables (Maris, 1995). For the probability that examinee i is engaged when attempting item j, $p(\Delta_{ij} = 1)$, we assume a Rasch model with

$$p(\Delta_{ij} = 1) = \frac{\exp(\phi_i - \iota_j)}{1 + \exp(\phi_i - \iota_j)}, \tag{9}$$

where ϕ_i denotes examinee i's engagement and ι_j gives item j's engagement difficulty. Examinee engagement determines whether examinees tend to approach items engagedly. Engagement difficulty determines how easily examinees interact with an item engagedly.

All person parameters are assumed to be multivariate normally distributed with mean vector

$$\boldsymbol{\mu}_{\mathcal{P}} = (\mu_\phi, \mu_\theta, \mu_\tau), \tag{10}$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\mathcal{P}} = \begin{pmatrix} \sigma_\phi^2 & \sigma_{\phi\theta} & \sigma_{\phi\tau} \\ \sigma_{\phi\theta} & \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\phi\tau} & \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}. \tag{11}$$

When a Rasch model is employed for responses and engagement indicators, the model can be identified by setting the expectations of ϕ, θ, and τ to zero. Item parameters are modelled as fixed effects.[2]

The proposed model's likelihood can be written as

$$\mathcal{L} = \prod_{i=1}^{N}\prod_{j=1}^{K}\Big(p(\Delta_{ij}=1|\phi_i,\iota_j)(1-d_{ij})p(u_{ij}|\theta_i,b_j)f(t_{ij}|\tau_i,\beta_j,\sigma_{\mathrm{E}}^2)+ \\ (1-p(\Delta_{ij}=1|\phi_i,\iota_j))p(d_{ij}|\gamma,\theta_i,\tau_i)p(u_{ij}|c)^{(1-d_{ij})}f(t_{ij}|\beta_D,\sigma_D^2)\Big)\cdot g(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\tau}|\boldsymbol{\mu}_{\mathcal{P}},\boldsymbol{\Sigma}_{\mathcal{P}}). \tag{12}$$

As can be seen, the framework allows for mixture distributions of responses and RTs at the item-by-examinee level, with the first row representing the model for engaged and the second the model for disengaged test-taking behaviour. The mixing proportions $p(\Delta_{ij}=1|\phi_i,\iota_j)$ and $1-p(\Delta_{ij}=1|\phi_i,\iota_j)$ are modelled as a function of examinee engagement ϕ_i and engagement difficulty parameters ι_j with an IRT model. $g(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\tau}|\boldsymbol{\mu}_{\mathcal{P}},\boldsymbol{\Sigma}_{\mathcal{P}})$ denotes the multivariate normal density of the person parameters. Note that in the case where examinee i omits item j, the first row does not contribute to the likelihood function, thereby incorporating the assumption that examinee i's engagement status is observable in the case where $d_{ij}=1$.

4. Prior distributions

Bayesian estimation techniques are employed to facilitate model estimation. For the prior distribution for the person parameter variance–covariance matrix $\boldsymbol{\Sigma}_{\mathcal{P}}$, we follow a separation strategy where the correlation matrix $\Omega_{\mathcal{P}}$ and person parameter standard deviations $\boldsymbol{S}_{\mathcal{P}}$ are separated out (Barnard, McCulloch, & Meng, 2000), that is,

$$\Sigma_{\mathcal{P}} = \mathrm{diag}(\boldsymbol{S}_{\mathcal{P}})\Omega_{\mathcal{P}}\mathrm{diag}(\boldsymbol{S}_{\mathcal{P}}). \tag{12}$$

Such separation strategies have been shown to yield unbiased parameter estimates of variances and correlations even under conditions with smaller sample sizes (Alvarez, Niemi, & Simpson, 2014). Furthermore, separation strategies circumvent the dependencies between variances and correlations inherent to inverse Wishart priors (Alvarez *et al.*, 2014; Gelman & Hill, 2007). Following recommendations by the Stan Development Team (2017), we employ an LKJ prior (Lewandowski, Kurowicka, & Joe, 2009) with shape 1 for the correlation matrix $\Omega_{\mathcal{P}}$, implying a uniform distribution on the correlation parameters and half Cauchy priors with location 0 and scale 5 for each element of $\boldsymbol{S}_{\mathcal{P}}$.

Following Fox (2010), we employ diffuse normal priors with mean 0 and standard deviation 10 for all engagement difficulties ι_j, difficulties b_j, time intensity offsets β_j^*, as well as the common mean β_{D} and each element of the vector of logistic regression parameters $\boldsymbol{\gamma}$. For residual standard deviations of logarithmized engaged RTs σ_{E} and the common standard deviation σ_{D} we suggest diffuse half Cauchy priors with location 0 and scale 5. For the common guessing parameter c we employ diffuse beta priors with $B(1,1)$.

[2] The model can easily be extended to assuming a joint distribution for item parameters. These are, however, challenging to estimate without bias under conditions with few items. Neglecting correlations of item parameters, in turn, has been shown not to affect unbiasedness and efficiency of parameter estimates (Molenaar, Oberski, Vermunt, & De Boeck, 2016; Molenaar, Tuerlinckx, & van der Maas, 2015).

5. Parameter recovery

To investigate estimability of the SA+E model, a simulation study was performed. We addressed two major research questions. First, the simulation study served to investigate whether true parameter values can satisfactorily be recovered under realistic conditions. Second, we aimed to identify boundary conditions concerning the sparseness of information on examinee disengagement for the detection thereof.

5.1. Data generation

Data were generated according to the SA+E model, employing R version 3.5.1 (R Development Core Team, 2017). To evaluate model performance under realistic research conditions, data-generating values were chosen to resemble parameter estimates reported in the empirical example below. To identify possible challenging conditions, we varied factors that are relevant for data sparseness in disengaged behaviour. Four variables were manipulated: the number of examinees (250, 500, 1,000), representing low, medium, and large sample sizes per item encountered in LSAs with balanced incomplete block designs (Gonzalez & Rutkowski, 2010); the number of items (10, 20); the rate of disengaged behaviour in the data set of size $N \times K$ (5%, 10%), reflecting rates of disengaged rapid guesses typically found in data from LSAs (Goldhammer *et al.*, 2016; Lee & Jia, 2014) as well as low to medium omission rates (OECD, 2013); and the percentage of omissions as opposed to guessing in disengaged behaviour (10%, 50%, 90%). Since omissions are assumed to occur only when examinees are disengaged, we suspect that sufficiently high omission rates facilitate estimation. Estimation might be more challenging when examinees mainly guess when disengaged, or when guessing is hard to detect due to low incidence.

Our manipulation of variables led to $3 \times 2 \times 2 \times 3 = 36$ conditions. For each condition, 50 data sets were generated. Using the MVRNORM function from the MASS package (Venables & Ripley, 2002), person parameters were randomly drawn from a multivariate normal distribution. We set engagement ϕ, ability θ, and speed τ variances to 3.50, 1.00, and 0.05, respectively. Correlations of engagement with ability, cor(ϕ,θ), and speed, cor(ϕ,τ), were set to .55 and .20, respectively. The correlation between ability and speed, cor(θ,τ), was set to $-.40$. Such negative correlations between ability θ and speed τ indicate that examinees showing higher levels of ability operate at a lower speed level and are rather common for low-stakes LSAs (Goldhammer *et al.*, 2014). For all item parameter types, we considered five different values, stemming from sequences $\{\iota_0 + 0.5l\}_{l=1}^{5}$ for engagement difficulties ι, $\{-1 + 0.5l\}_{l=1}^{5}$ for difficulties b, and $\{3 + 0.25l\}_{l=1}^{5}$ for time intensities β. For tests of length $K = 10$ and $K = 20$ these sequences were repeated twice and four times, respectively. To obtain rates of disengaged behaviour of 5% and 10%, ι_0 was set to -5 and -4.25, respectively. This resulted in item-level disengagement rates ranging from 1.11% to 8.20% and from 2.35% to 17.38% under conditions with overall disengagement rates of 5% and 10%, respectively. The logistic regression parameters were set to $\gamma_\theta = -1$ and $\gamma_\tau = -10$. For omission rates in disengaged behaviour of 10%, 50%, and 90%, the intercept was set to $\gamma_0 = -3$, $\gamma_0 = 0$, and $\gamma_0 = 3$, respectively. The probability correct for disengaged responses was set to $c = .25$ for all items. Logarithmized disengaged RTs were drawn from a normal distribution with mean $\beta_D = 3$ and variance $\sigma_D^2 = 1.95$. The common residual variance for logarithmized engaged RTs was set to $\sigma_D^2 = 0.15$.

5.2. Estimation procedure

Bayesian estimation was conducted using Stan version 2.18 (Carpenter *et al.*, 2017), employing the RSTAN package (Guo, Gabry, & Goodrich, 2018) for R version 3.5.1 (R Development Core Team, 2017). For sampling from the posterior distributions, Stan employs the No-U-Turn sampler (Hoffman & Gelman, 2014), an adaptive form of Hamiltonian Monte Carlo sampling (Neal, 2011). Data were analysed employing the SA+E model. On each data set, we ran four Markov chain Monte Carlo (MCMC) chains with 10,000 iterations each, with the first 5,000 employed as warm-up. The number of iterations was chosen based on conclusions drawn from pre-analyses, inspecting potential scale reduction factor (PSRF) values, trace plots, and effective sample sizes (ESSs). Stan code for the SA+E model is provided in the Supporting Information.

5.3. Results

Statistical performance was evaluated in terms of convergence and efficiency of the MCMC chains as well as bias and efficiency of parameter estimates. We assessed convergence on the basis of PSRF values. Replications with PSRF values below 1.10 for all parameters were considered as being converged (Gelman & Rubin, 1992; Gelman & Shirley, 2011). The efficiency of the estimation procedure was evaluated by considering ESS (Kass, Carlin, Gelman, & Neal, 1998), indicating the degree of precision with which the empirical mean of the MCMC chains approximates the expected value of the posterior distribution (Lüdtke, Robitzsch, & Wagner, 2018). Following Zitzmann and Hecht (2019), we considered an ESS above 400 for all parameters as sufficient.

Table 1 displays proportions of replications with PSRF values below 1.10 as well as ESSs above 400 across all conditions. Convergence rates as indicated by PSRF values below 1.10 were at least 90% under all conditions with $K = 20$ items. Under conditions with $K = 10$ and smaller sample sizes ($N \leq 500$), however, convergence was challenged, with the lowest convergence rate being .82.

In some cells of the simulation design with $N \leq 500$, proportions of replications with ESSs for all parameters higher than 400 were somewhat lower than convergence rates as evaluated on the basis of PSRF values. This indicates that although the chains converged and mixed well, the parameter space was explored rather slowly and more iterations might be needed to ensure good approximation of the posterior mean (Zitzmann & Hecht, 2019). Further assessments of convergence behaviour of replications with PSRF values above 1.10 showed very poor, if any, mixing of the MCMC chains, with PSRF values of up to 573.45, indicating that engaged and disengaged behaviours were not separable. Since the mean across chains that did not show any mixing is not meaningful and non-coverged solutions would not be interpreted in practice, we excluded replications with PSRF values exceeding 1.10 from all subsequent analyses.

To evaluate bias and efficiency of parameter estimates, we assessed the median and 50% ranges of posterior means. Good parameter recovery was found under all conditions with a sufficiently high number of examinees ($N = 1{,}000$) and items ($K = 20$). Under conditions with fewer items and examinees, engagement variance, engagement difficulty, as well as regression parameters for predicting the probability of omitting rather than guessing when being disengaged were sensitive to bias when little information on examinee disengagement was available. All remaining parameters could be recovered without systematic bias across all conditions of the simulation design. Results for these are given in the Supporting Information. As was to be expected, efficiency in parameter estimates as indicated by narrower 50% ranges increased with an increasing number of

Table 1. Proportions of replications with PSRF values < 1.10 and ESS > 400 for all parameters after 10,000 iterations

N	*K*	Disengaged (%)	Omitted (%)	PSRF < 1.10	ESS > 400
250	10	5	10	1.00	.94
			50	.84	.74
			90	.96	.96
		10	10	.92	.82
			50	.86	.84
			90	.92	.92
	20	5	10	.96	.96
			50	1.00	1.00
			90	.96	.96
		10	10	1.00	1.00
			50	.98	.98
			90	1.00	1.00
500	10	5	10	.92	.88
			50	.98	.94
			90	.94	.94
		10	10	.96	.96
			50	.98	.98
			90	.82	.82
	20	5	10	1.00	1.00
			50	.98	.98
			90	.94	.94
		10	10	1.00	1.00
			50	.98	.98
			90	.94	.92
1,000	10	5	10	1.00	1.00
			50	.96	.96
			90	.88	.86
		10	10	.98	.98
			50	.98	.98
			90	.92	.92
	20	5	10	.98	.98
			50	1.00	1.00
			90	1.00	1.00
		10	10	.98	.98
			50	.98	.98
			90	.98	.98

Note. Omissions give the percentage of item omissions on disengaged behaviour.
N = number of examinees; K = number of items.

both examinees and items for all parameter types. In addition, parameters associated with disengagement were estimated more efficiently under conditions with higher omission rates, that is, under conditions with a higher portion of disengaged behaviour being directly observable.

Results for person parameter variances and correlations are given in Figure 2. Engagement variance var(ϕ) estimates were upwardly biased under conditions with sparse information on examinee disengagement, such that under rather challenging conditions with only 250 examinees, 10 items, and a low disengagement rate of 5% out of

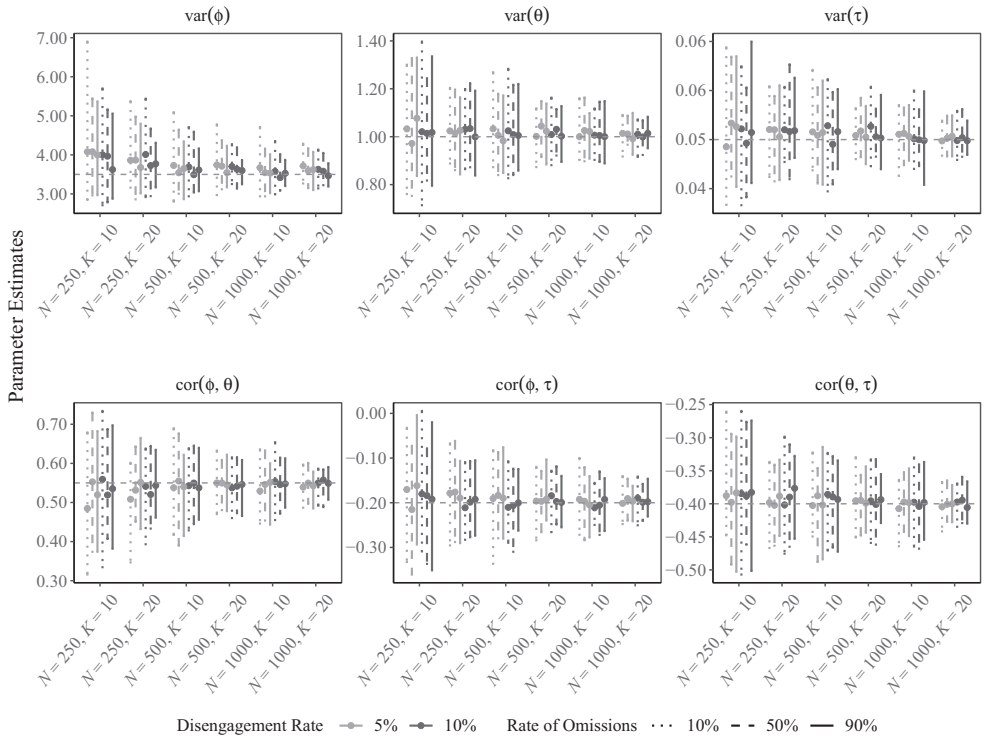

Figure 2. Medians and 50% ranges of person parameter variance and correlation estimates. The dashed horizontal line indicates the respective true parameter. Note that *y*-axes differ in scale. ϕ = engagement; θ = ability; τ = speed; N = number of examinees; K = number of items. The shades of the lines denote the rates of disengaged behaviour. The percentages of omissions on disengaged behaviour are given by different line types.

which only 10% went back to item omissions, median var(ϕ) estimates of 4.09 were observed, as compared to the data-generating value of 3.50. However, bias decreased rapidly with an increasing number of examinees as well as higher omission rates, such that under conditions with $N \geq 500$, medians of posterior means were extremely close to the true value.

Likewise, under conditions with smaller data sets, engagement difficulties ι were sensitive to bias for items with low rates of disengaged behaviour, that is, when the true parameter was small (see Figure 3). This effect was further intensified when disengaged behaviour was not directly observable and consisted predominantly of random guesses. Accordingly, parameter estimates for the smallest data-generating value assessed in the simulation study of -4.50 (corresponding to an item-level disengagement rate of only 1%) were most sensitive to bias under conditions with only 10% of disengaged behaviour resulting in item omissions, such that under the condition with only $N = 250$ examinees and $K = 10$ items, a median of parameter estimates of -4.74 was observed. Bias decreased rapidly with an increasing number of examinees as well as higher percentages of omissions on disengaged behaviour. Under conditions with $N = 1{,}000$ examinees, differences were extremely close to zero for all values of ι considered.

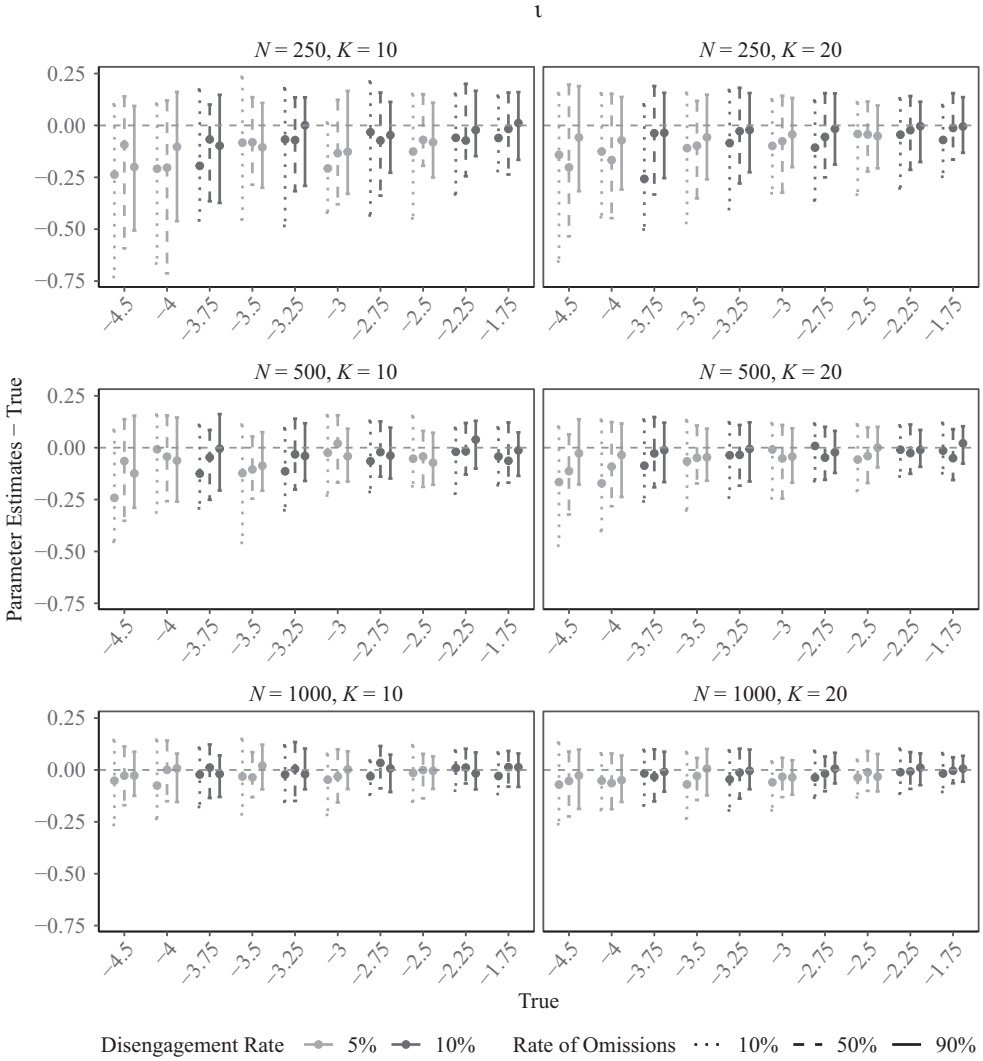

Figure 3. Medians and 50% ranges of differences between estimated and true engagement difficulties ι plotted against the true parameters. The dashed horizontal line indicates a difference of zero. N = number of examinees; K = number of items. The shades of the lines denote the rates of disengaged behaviour. Different percentages of omissions on disengaged behaviour are given by different line types.

Regression parameters were challenging to estimate under conditions with less than $K = 20$ items (see Figure 4). Under such conditions, highly negative as well as highly positive intercepts γ_0, resulting in disengaged behaviour consisting mainly of rapid guesses and omissions, respectively, were biased with median parameter estimates ranging from −2.95 to −3.47 and from 3.16 to 3.76, as compared to the true values of −3 and 3, respectively. In addition, slopes for the regression of omission probability on speed, γ_τ, were underestimated under conditions with $K = 10$.

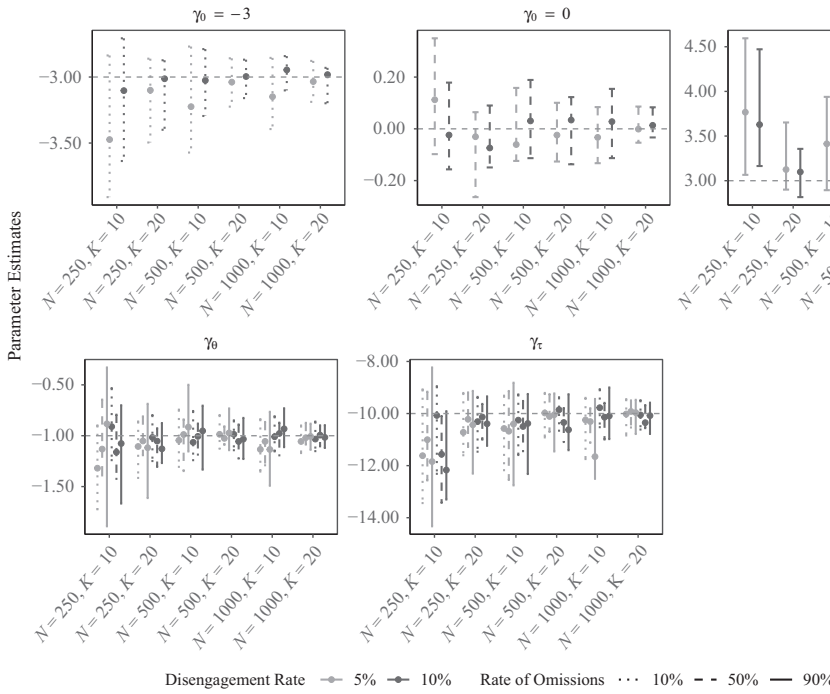

Figure 4. Medians and 50% ranges of regression parameters. The dashed horizontal line indicates the respective true parameter. Note that *y*-axes differ in scale. γ_0 = intercept; γ_θ = slope for regression of omission probability on ability; γ_τ = slope for regression of omission probability on speed; N = number of examinees; K = number of items. The shades of the lines denote the rates of disengaged behaviour. Different percentages of omissions on disengaged behaviour are given by different line types. Note that differences in omission rates were induced by different values for γ_0.

6. Illustrating the model

To illustrate how the SA+E model differs conceptually from current approaches for identifying examinee disengagement as well as for handling item omissions, we took data from single replications of the simulation study for conditions with a disengagement rate of 10% out of which 50% went back to item omissions, and compared parameter estimates for the SA+E model with those obtained from models that either model the occurrence of item omissions but assume all observed responses to stem from engaged response processes, or filter disengaged behaviour but assume engagement to be unrelated to ability and item omissions to be ignorable. We chose to compare the SA+E model to the speed-accuracy + omission (SA+O) model (Ulitzsch *et al.*, 2019) and the mixture model for identifying examinee engagement presented by Wang and Xu (2015), representing two recent modelling approaches for omissions and the identification of disengaged guessing behaviour, respectively. Adopting the graphical notation of the SA+E framework, the models are depicted in Figures 5 and 6, respectively.

The SA+O model models the omission process according to equation (2). All responses are assumed to stem from engaged response processes and thus modelled with a Rasch model as in equation (1). Different data-generating processes are assumed for RTs associated with responses and omission, respectively. RTs associated with responses are modelled as a function of speed and time intensity, as in equation (3). RTs associated with

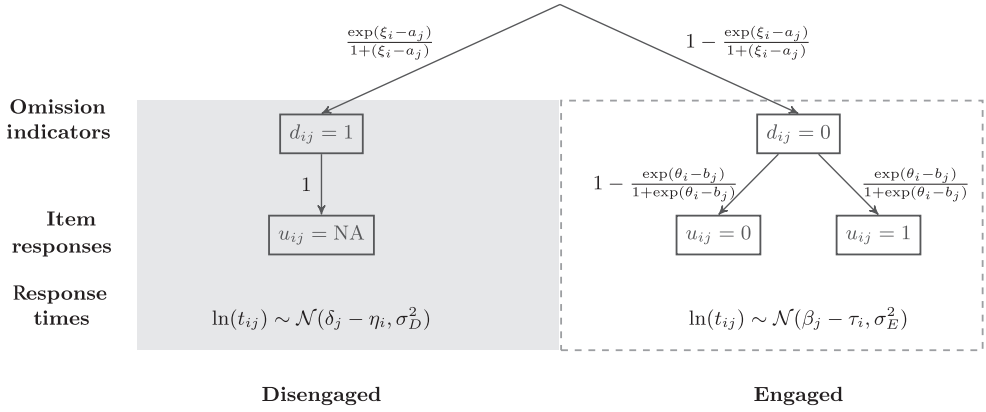

Figure 5. SA+O model by Ulitzsch et al. (2019).

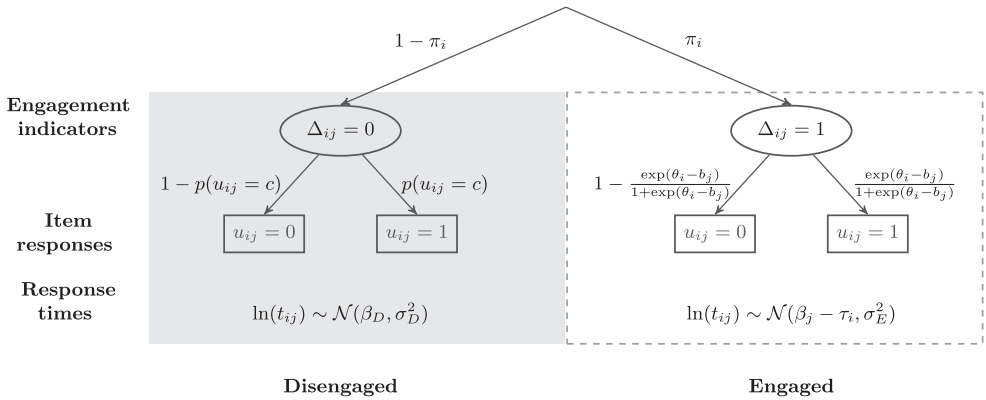

Figure 6. Mixture model for identifying examinee engagement by Wang and Xu (2015).

item omission are modelled analogously, but with a different set of item and person parameters (omission time intensity and omission speed) thereby allowing for examinees to operate at different speed levels when generating responses and omitting. The SA+O model assumes a joint distribution for ability, speed, omission propensity, and omission speed.

The Wang and Xu model is a mixture modelling approach for examinee disengagement in terms of guessing. The model assumes a person-specific disengagement probability that is constant across items and distinct from ability, that is, $p(\Delta_{ij} = 1) = \pi_i$, with π_i denoting examinee i's engagement probability. For responses and RTs, the model assumes models for engaged and disengaged examinees that are equivalent to those assumed in the SA+E framework. When specifying the Wang and Xu model, item omissions were ignored. The Wang and Xu model differs from the SA+E model in the treatment of item omissions as well as in that it assumes engagement probability to be unrelated to ability and constant across items.[3] All models were estimated employing the same set-up for model estimation as in the simulation study.

[3] For studies on assessing these issues separately, see Pokropek (2016), Pohl et al. (2014), and Wang and Xu (2015).

To investigate the effects of different test lengths and sample sizes, we varied the number of items (10, 20) and examinees (250, 500, 1,000).

Differences in ability estimates between the SA+O model and the SA+E model (given in Figure 7 as a function of engagement estimated using the SA+E framework as well as the number of item omissions) are close to zero for examinees with high engagement, that is, for examinees who rarely guess or omit items. With increasing disengagement, however, there are increasing differences in ability estimates between the SA+O and SA+E models. This goes back to assuming all responses to be engaged as well as misspecifying engagement (or omission propensity) by neglecting the fact that disengaged examinees tend not only to omit but also to guess. This is also reflected in the differences in item difficulties (given in Figure 8 as a function of engagement difficulty). Due to assuming all responses to be engaged, difficulties of easy items tend to be overestimated. This effect is especially pronounced for items with higher engagement difficulties, as these tend to be guessed on more often.

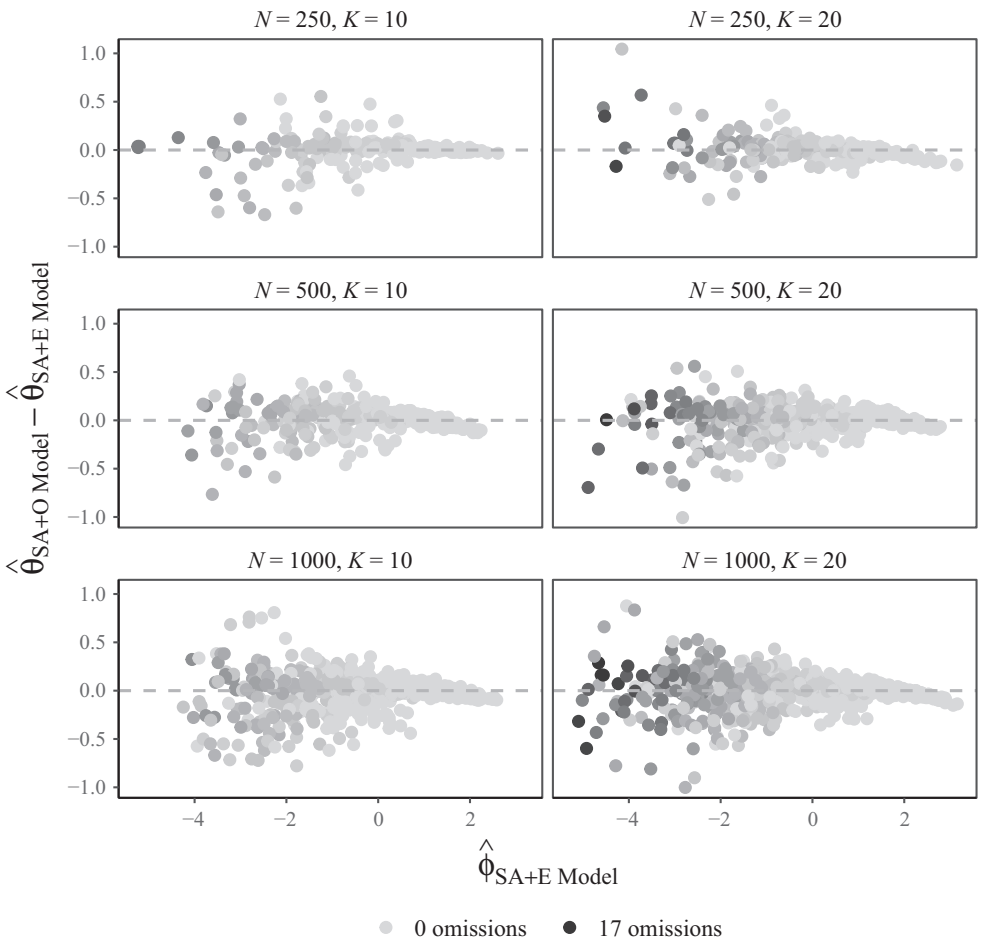

Figure 7. Differences in ability estimates retrieved from the SA+O model and the SA+E model plotted against engagement estimates retrieved from the SA+E model. The colour of the points denotes the number of item omissions for each examinee, with darker shades indicating a higher number of item omissions. N = number of examinees; K = number of items.

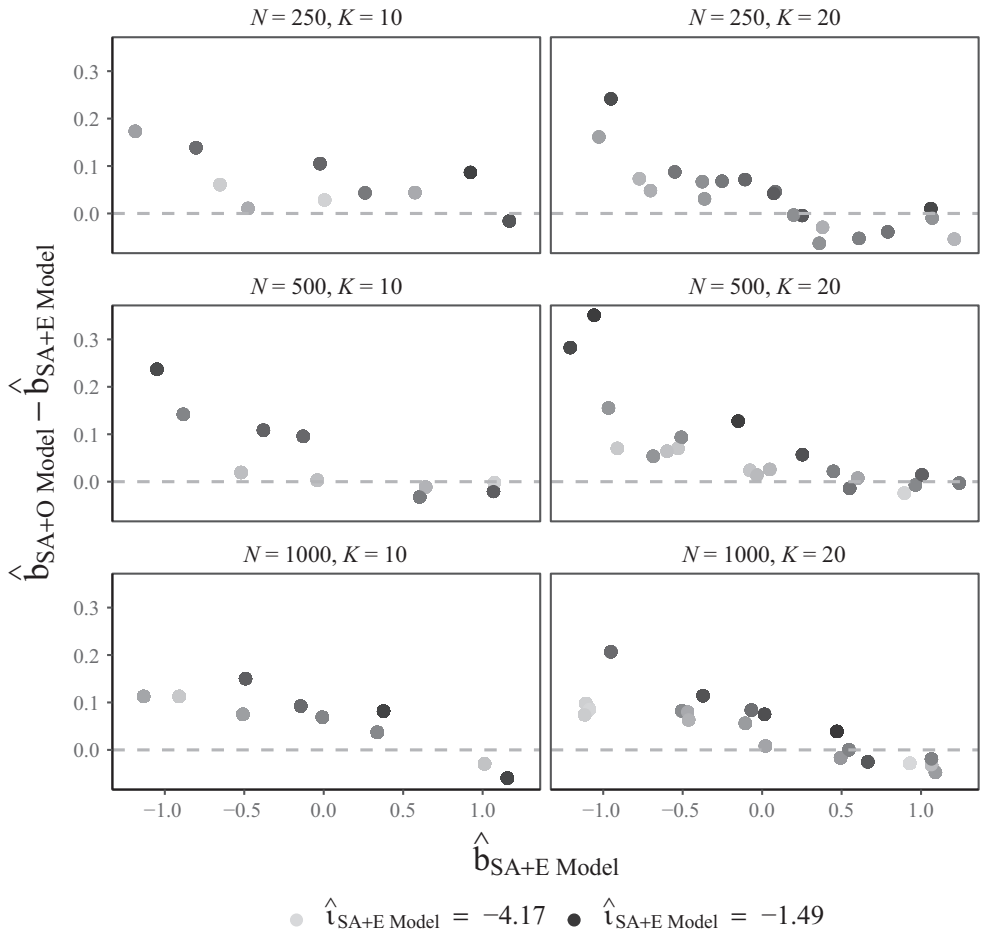

Figure 8. Differences in item difficulty estimates retrieved from the SA+O model and the SA+E model plotted against item difficulty estimates retrieved from the SA+E model. The colour of the points denotes the items engagement difficulty estimates retrieved from the SA+E model, with darker shades indicating higher engagement difficulty. N = number of examinees; K = number of items.

The Wang and Xu model, too, gives ability estimates for examinees with higher engagement that are very close to those obtained from the SA+E model (see Figure 9). However, due to neglecting the fact that ability and engagement are positively related, ability for examinees with lower engagement is overestimated by the Wang and Xu model. This also results in systematically lower item difficulties (see Figure 10), with differences being higher for easy items. Since in the data-generating model ability and engagement are positively correlated, observed engaged responses are more likely to be observed for more able examinees, resulting in difficulties being underestimated (Rose, 2013).

For both model comparisons, differences in ability and item difficulty estimates are similar for different numbers of examinees and items.

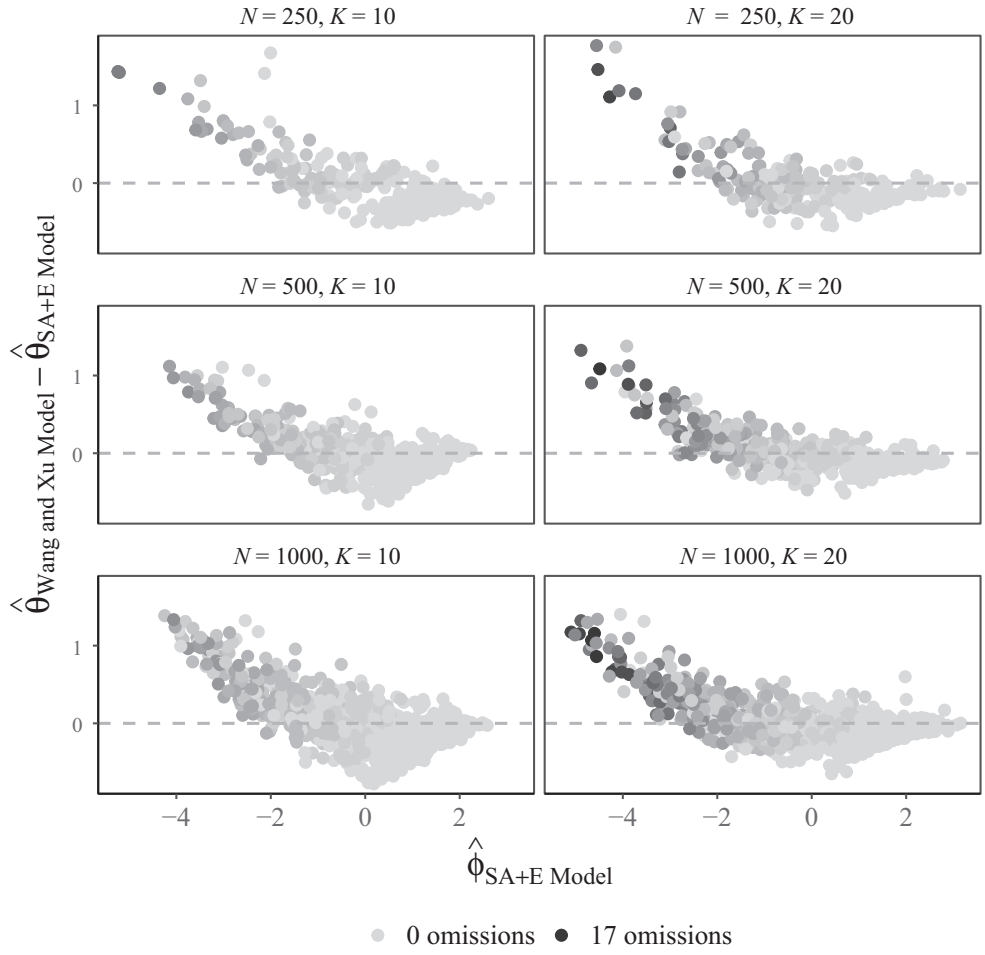

Figure 9. Differences in ability estimates retrieved from the Wang and Xu model and the SA+E model plotted against engagement estimates retrieved from the SA+E model. The colour of the points denotes the number of item omissions for each examinee, with darker shades indicating a higher number of item omissions. N = number of examinees; K = number of items.

7. Empirical example

To illustrate the use of the SA+E model for detecting and understanding disengagement, we employed data from PISA 2015. We focused on mathematical literacy block number 1, comprising $K = 12$ items, out of which three had an OR format and nine were MC. For reasons of simplicity, we dichotomized partial credit items, scoring partially correct as incorrect. We applied the model to several samples of students from different countries, all of which led to comparable conclusions. Exemplarily, results for the Austrian subset, containing $N = 844$ examinees, are reported. The data set under consideration had an omission rate of 10.40%. Item-level omission rates ranged from 0.04% for the MC item administered at position 1 to 34.60% for the OR item administered at position 5. An additional 0.48% of responses were missing due to not-reached items. These were ignored in the estimation.

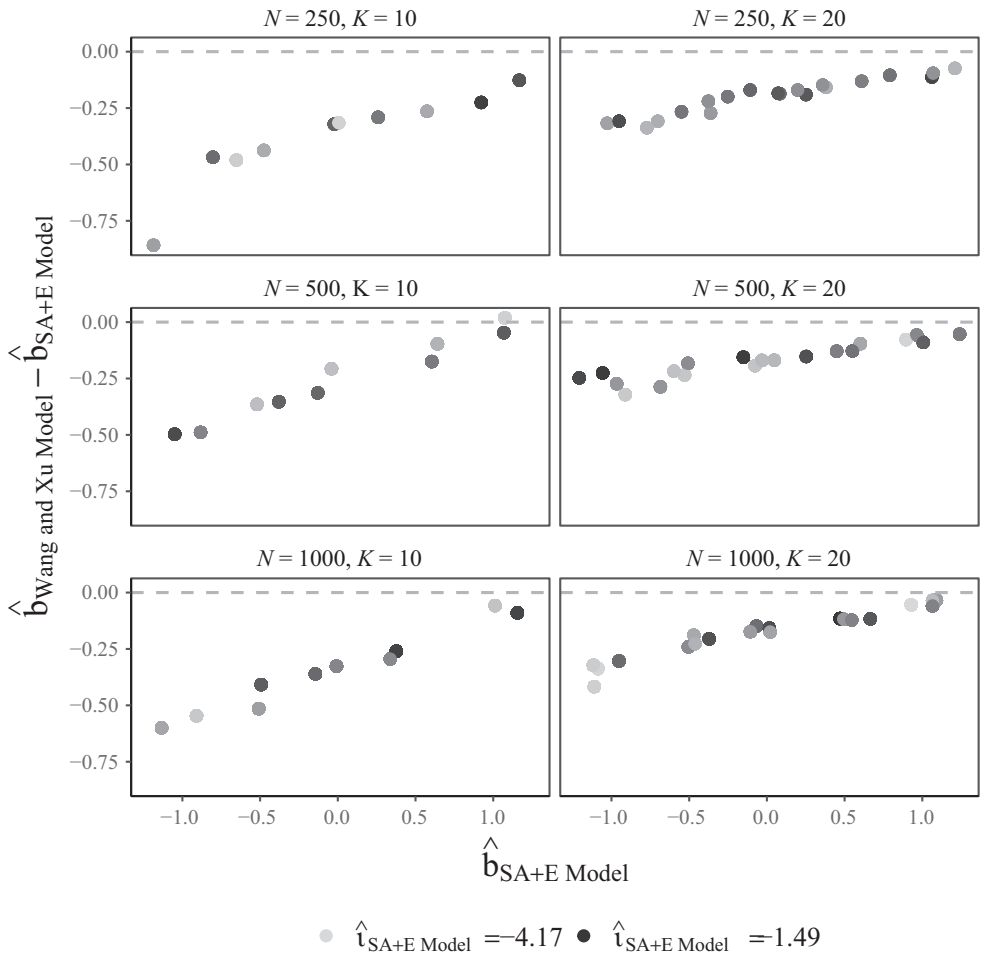

Figure 10. Differences in item difficulty estimates retrieved from the Wang and Xu model and the SA+E model plotted against item difficulty estimates retrieved from the SA+E model. The colour of the points denotes the items engagement difficulty estimates retrieved from the SA+E model, with darker shades indicating higher engagement difficulty. N = number of examinees; K = number of items.

7.1. Estimation and model checking

For estimation, the same set-up as in the simulation study was employed. To take into account that the item block contained different item types, we specified item-type-specific probabilities correct when answering perfunctorily (c_O) or guessing (c_M) on OR and MC items, respectively. In addition, we allowed for item-type-specific regression intercepts γ_{O0} and γ_{M0} determining the probability of omitting instead of perfunctorily answering or guessing on an item with an OR or MC format, respectively.

After 10,000 iterations per chain, the highest PSRF value and lowest ESS were 1.002 and 3,471.55, respectively. Model fit was evaluated employing posterior predictive checks (Gelman & Hill, 2007). For these, we simulated 30 data sets by drawing parameters from the posterior distribution and visually compared observed and simulated proportions correct and omitted as well as distributions of observed and simulated RTs. RT

distributions were predicted well by the model. Although overall proportions correct were predicted well by the model, for some items, comparisons of observed and predicted probabilities correct indicate that a more complex measurement model, such as a two-parameter logistic model, might fit the data better. Likewise, for some items, the model underpredicted item omissions for examinees with higher proportions of item omissions. Possible model extensions with less restrictive assumptions concerning measurement models for responses, RTs, and latent response indicators are addressed in the discussion in Section 8. Plots for posterior predictive checks are given in the Supporting Information.

7.2. Results

Probabilities correct for disengaged guesses on MC items and perfunctory answers on OR items indicate that while examinees correctly guessed with a probability of .23 [.19, .28] on MC items, it was highly unlikely (.11 [.08, .15]) to answer correctly on an OR item when answering only perfunctorily. Examinees tended to spend on average $\exp(\beta_D) = \exp(3.54) = 34.53$ s on an item when approaching it in a disengaged manner. At the same time, there was considerable variation in logarithmized RTs associated with disengaged behaviour, with $\sigma_D^2 = 1.40$ [1.31, 1.49]. Means of the posterior distribution of person parameter variances and correlations, together with 95% highest density intervals, are displayed in Table 2. More able examinees tended to be more engaged. Furthermore, engaged as well as more able examinees tended to work at a slower pace when generating engaged responses. The intercepts of the logistic regression predicting the probability of omitting rather than randomly guessing or answering perfunctorily indicate that examinees with average ability and speed were more likely to guess and less likely to perfunctorily answer than to omit ($\gamma_{M0} = -0.71$ [−0.94, −0.47], $\gamma_{O0} = 0.45$ [0.26, 0.67]). The slopes indicate that examinees with higher ability and higher speed tended to guess or perfunctorily answer rather than to omit when disengaged ($\gamma_\theta = -0.74$ [−0.98, −0.52]; $\gamma_\tau = -4.79$ [−6.04, −3.70]). Item parameters and 95% highest density intervals are depicted in Figure 11. Item numbers for OR items are given in bold type. Examinees were more likely to disengage on more difficult items ($\mathrm{cor}(\iota, b) = .68$) as well as on items with higher time intensity offsets ($\mathrm{cor}(\iota, \beta^*) = .81$). Time intensity offsets β^* indicate that examinees tended, on average, to require $\exp(0.24) = 1.27$ to $\exp(1.73) = 5.65$ times longer to generate engaged responses to these items than they tended to interact with items in a disengaged manner. Engagement difficulty parameters ranged from −6.55 (item 1) to 0.60 (item 5). For these, respectively 0.05% and 59.22% of item-by-examinee interactions were classified as disengaged. Note that for item 5, the model-implied disengagement rate was notably higher than the item-level omission rate of 34.60%. The difference between the expected engagement rate and the observed omission rate can be attributed to guessing (or perfunctory answers), illustrating that examinees both omitted

Table 2. Person parameter variances and correlations

	ϕ	θ	τ
ϕ	3.25 [2.65, 3.93]		
θ	.59 [.50, .68]	1.47 [1.23, 1.74]	
τ	−.35 [−.44, −.25]	−.36 [−.45, −.26]	0.04 [0.03, 0.05]

Note. Highest density intervals are given in square brackets.
ϕ = engagement; θ = ability; τ = speed.

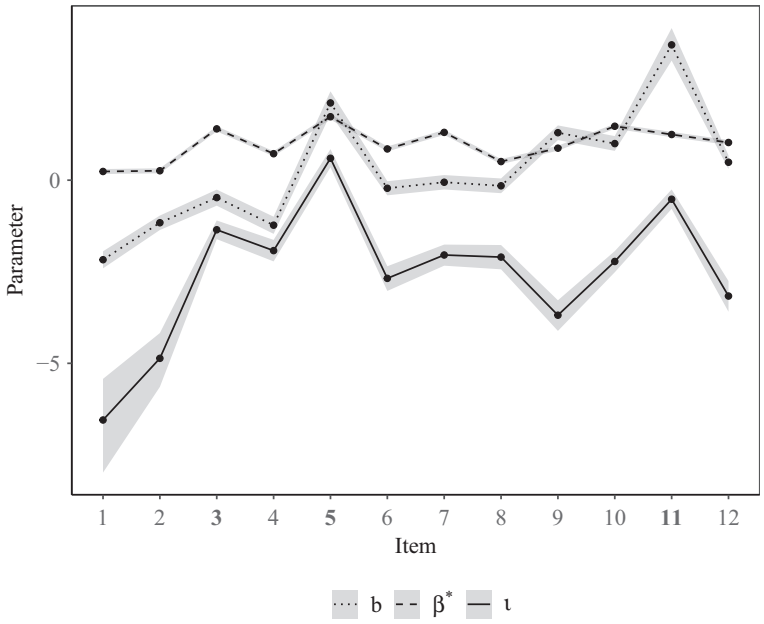

Figure 11. Posterior means and highest density intervals for engagement difficulties ι, difficulties b, and time intensity offsets β^*. Items in bold have an open -response format.

and answered perfunctorily when disengaged. This is also illustrated in an overall expected disengagement rate of 18.23% as compared to the omission rate of 10.40%.

8. Discussion

The SA+E model presented in this paper brings together research on modelling examinee engagement and research on missing values and provides a framework for identifying and modelling examinee disengagement in terms of both random guesses and perfunctory answers as well as in terms of omissions. By employing a latent response approach with engagement probabilities modelled as a function of person and item parameters, the model allows for classifying disengaged behaviour at the item-by-examinee level as well as for assessment of item and examinee characteristics associated with such behaviour. In addition, the model allows for differences in disengaged test-taking behaviour across examinees by regressing the probability of omitting rather than randomly guessing or answering perfunctorily on ability and speed.

The SA+E framework complements and refines recent approaches for examinee disengagement as well as non-ignorable item omissions. Compared to RT-based scoring methods separating engaged and disengaged responses and/or item omissions by defining RT thresholds (Frey *et al.*, 2018; Lee & Jia, 2014; Wise & DeMars, 2006), the SA+E framework comes with less strict assumptions concerning RT distributions associated with engaged and disengaged behaviour since these are allowed to overlap. Compared to previous model-based approaches for identifying disengaged examinee behaviour (Meyer, 2010; Pokropek, 2016; Schnipke & Scrams, 1997; Wang & Xu, 2015), the model allows disengaged behaviour to vary across both items and examinees while considering engagement when estimating ability.

In this regard, the model also adds to a broader class of models that employ mixture modelling to identify differences in examinee behaviour. Few model-based approaches for detecting differences in examinee behaviour allow for differences at the item-by-examinee level (Erosheva, 2002; Molenaar & de Boeck, 2018; Pokropek, 2016). These do, however, not allow these differences to be related to different levels of ability. In this context, the proposed framework can be adapted to suit other applications and further model developments seeking to identify behavioural differences at the item-by-examinee level while modelling the underlying processes jointly with ability.

We illustrated the model's advantages by showing that ability estimates and item difficulties can be biased when neglecting the fact that examinees tend to omit and guess when disengaged, that engagement is related to ability, and that engagement probabilities tend to vary across items. Our findings corroborate findings from previous studies on ignoring guessing behaviour and item omissions as well as on neglecting the relationship between engagement and ability (Pohl *et al.*, 2014; Pokropek, 2016; Rios *et al.*, 2017; Rose *et al.*, 2010; Wang & Xu, 2015).

The model yields unbiased and efficient parameter estimates under conditions with at least $N = 500$ examinees and $K = 20$ items even under disengagement rates of as low as 5% and unbalanced proportions of item omissions and guesses for disengaged behaviour. Under conditions with fewer items or examinees, low disengagement rates pose a threat to obtaining unbiased and efficient parameter estimates. We therefore recommend applying the model to smaller data sets with $N < 500$ or $K < 20$ only when omission rates are high, that is, at least 5%. Due to the model's complexity, convergence might be more challenging to achieve under conditions with few items and examinees.

When no convergence can be reached it is likely that disengaged behaviour predominantly consists of item omissions (e.g., for tests with complex item formats where observed responses are unlikely to go back to guesses or perfunctory answers) and model-based approaches for modelling omission processes (Holman & Glas, 2005; Ulitzsch *et al.*, 2019) pose a less complex alternative to the SA+E model. When seeing item omissions as indicators of disengaged behaviour, omission propensity in model-based approaches for item omissions is equivalent to the engagement variable, with examinee disengagement manifesting itself only in item omissions, while all observed responses are assumed to stem from engaged response processes. Under such assumptions, examinee engagement would be fully observable, with engagement indicators Δ_{ij} corresponding to the negation of omission indicators $1 - d_{ij}$ (see the Supporting Information). Likewise, when no omissions occurred, the model can easily be simplified to assuming disengaged behaviour to a result in guessing only while still jointly modelling engagement, ability, and speed (see the Supporting Information).

In the empirical example we found examinee engagement and ability to be related. At the same time, the only moderate correlation between engagement and ability provides supporting evidence that engagement and ability represent different constructs. In addition, we found engagement to vary largely across items and examinees. Items that were more complex in terms of difficulty and time intensity were found to evoke disengagement more easily. This is in line with findings from previous studies employing threshold methods for identifying examinee disengagement (Lee & Jia, 2014; Wise *et al.*, 2009). In addition, we illustrated that both item omissions and guessing are prevalent in LSA data and thus both need to be considered.

8.1. Limitations and future directions

In the SA+E model, identifying examinee disengagement is facilitated by assuming that all item omissions stem from examinee disengagement while allowing for observed responses to stem from either solution or guessing behaviour. Thus, similarly to previous model-based approaches for item omissions (Holman & Glas, 2005; Ulitzsch *et al.*, 2019), the SA+E model assumes that all item omissions stem from the same data-generating process. The model does not allow for engaged item omissions, which might occur when examinees omit items after seriously reading and considering the item. Such mechanisms have been discussed to be plausible (Becker & Pohl, 2016; Mislevy & Wu, 1996; Robitzsch, 2014). Extending the model to allow for different omission mechanisms is therefore a pertinent topic for future research.

Furthermore, examinees might not work with a constant level of engagement throughout the test. While some examinees might be disengaged throughout the whole test, it is easy to imagine that others might be engaged at the beginning of the assessment but become more disengaged towards the end. Non-stationarity of person variables can in principle be incorporated by adding additional linear or nonlinear terms (see Fox & Marianti, 2016, for an extension of the speed-accuracy model that allows for varying speed across the test).

Although the SA+E model allows for the occurrence of item omissions to vary across examinees when these approach an item in a disengaged manner, it is still rather restrictive in that it assumes the probability of omitting rather than guessing to be a function of ability and speed. A variety of other examinee- or item-specific factors such as demographic variables or item features might determine disengaged test-taking strategies. Considering these therefore constitutes a promising extension of the SA+E model.

The proposed model assumes examinee disengagement to result in random guesses, perfunctory answers, and item omissions. Examinee disengagement can, however, manifest itself in a variety of test-taking behaviours different from those considered in the proposed model. Examinees could, for instance, still employ solution strategies on an item but just try less hard (Debeer & Janssen, 2013) or still use their ability to some extent for differentiating among responses while guessing (San Martín, del Pino, & de Boeck, 2006). In its most extreme form examinee disengagement might result in quitting the assessment altogether. In fact, examinees who spend only a short time on a test without reaching the time limit or the end of the test are more likely to guess on the items they attempted (Cao & Stokes, 2008). A model for not-reached items due to quitting has been proposed by Ulitzsch, von Davier, and Pohl (in press). Integrating research on modelling quitting behaviour with research on examinee disengagement would enrich research on examinee disengagement as well as provide further insights into examinee test-taking behaviour.

Assessing the joint distribution of person variables yields valuable insights into examinee behaviour. In addition, relating engagement to, for example, demographic variables or personality can provide additional insight into possible reasons for examinee disengagement or for identifying groups of persons with a high prevalence of disengagement. For instance, omission propensity has been shown to be relatively stable across different domains and to be related to demographic variables such as gender (Köhler *et al.*, 2015a). Similar effects could be expected for examinee engagement. Furthermore, relating examinee engagement to self-reported test-taking motivation, as for example administered in PISA (OECD, 2017), could be used to validate the assumptions made in the proposed model.

The model was presented employing a Rasch model for item responses as well as its RT equivalent, with time discrimination parameters fixed to be the same across

all items. Although Rasch modelling is in accordance with the analysis frameworks of major LSAs (OECD, 2017; Pohl & Carstensen, 2012), such assumptions might not always hold for the data at hand. In fact, in the empirical example, posterior predictive analyses revealed that less restrictive measurement models for responses might indeed fit the data better. Implementing these into the model, however, is not trivial since it can be challenging to distinguish engaged and disengaged responses on items with low item discrimination. Such model extensions therefore remain a task for future research.

Similarly, in future research, the model could be extended to include more complex measurement models for latent response indicators. For these, the SA+E model assumes a unidimensional Rasch model. This assumption is likely to be violated when examinees differ in the level of engagement with which they approach different types of items. Previous research suggests that this might indeed be the case. Omission behaviour, for instance, has often been found to differ for items with a simple MC format and items with a more complex response format (Köhler *et al*., 2015b; Koretz, 1993). Likewise, in the empirical example, for some items, the model underpredicted item omissions for examinees with a higher number of omissions. In this context, specifying a multidimensional measurement model for latent response indicators might model disengaged test-taking behaviour more adequately.

In addition, Molenaar, Bolsinova, and Vermunt (2018) have shown that violations of the lognormal assumptions for RTs may jeopardize correct classifications in mixture IRT models employing RTs for identifying differences in examinee behaviour. As a solution, Molenaar *et al.* (2018) suggested a semi-parametric approach based on categorizing RTs that can easily be integrated with the SA+E framework.

In the current paper, Bayesian techniques were employed for model estimation. Although this yielded good parameter recovery with sample sizes of as low as $N = 500$, estimation was rather time-intense: under the conditions with the largest data sets ($N = 1{,}000$, $K = 20$), estimation took approximately 24 hr. Research questions in educational research often concern multiple groups, constructs, or points in time, and involve larger data sets. Bayesian estimation might thus not always be feasible. With technical and algorithmic advances, we expect this to be resolved. Until then, future research may also consider the feasibility of maximum likelihood estimation for the proposed model. Here the challenge will be to obtain convergence and valid solutions when the prevalence of item omissions and guessing is low on some items.

Acknowledgements

This work was supported by the German Research Foundation (DFG) under Grant PO 1655/2-2.

References

Alvarez, I., Niemi, J., & Simpson, M. (2014). *Bayesian inference for a covariance matrix*. Preprint, arXiv:1408.4050v2.

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*(4), 1281–1311.

Becker, B. & Pohl, S. (2016). *Missing values in IRT-analyses of large-scale assessments: Do model-based approaches perform well under different underlying missing values processes?*. Paper presented at VII European Congress of Methodology, Palma, Spain.

Bhola, D. S. (1994). *An investigation to determine whether an algorithm based on response latencies and number of words can be used in a prescribed manner to reduce measurement error* (Doctoral dissertation). University of Nebraska-Lincoln. Retrieved from https://search.proquest.com/docview/304127164

Boe, E. E., May, H., & Boruch, R. F. (2002). *Student task persistence in the third international mathematics and science study: A major source of achievement differences at the national, classroom, and student levels* (Techical Report No. CRESP-RR-2002-TIMSS1). Philadelphia, PA: Center for Research and Evaluation in Social Policy, University of Pennsylvania.

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record*, *113*(11), 2309–2344.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*, 209–230. https://doi.org/10.1007/S11336-007-9045-9

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Cosgrove, J. (2011). *Does student engagement explain performance on PISA? Comparisons of response patterns on the PISA tests across time*. Dublin, Ireland: Educational Research Centre. Retrieved from http://www.erc.ie/documents/engagement_and_performance_over_time.pdf

Culbertson, M. J. (2011, April). *Is it wrong? Handling missing responses in IRT*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

de Ayala, R., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, *38*(3), 213–234. https://doi.org/10.1111/j.1745-3984.2001.tb01124.x

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164–185. https://doi.org/10.1111/jedm.12009

Erosheva, E. A. (2002). *Grade of membership and latent structure models with application to disability survey data* (Doctoral dissertation). Carnegie Mellon University. Retrieved from https://pdfs.semanticscholar.org/1fe4/64b6cae48d009697783bdbb72bcd4527608a.pdf

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225–245. https://doi.org/10.1111/j.1745-3984.2008.00062.x

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York: 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. https://doi.org/10.1080/00273171.2016.1171128

Frey, A., Spoden, C., Goldhammer, F., & Wenzel, S. F. C. (2018). Response time-based treatment of omitted responses in computer-based testing. *Behaviormetrika*, *45*(2), 505–526. https://doi.org/10.1007/s41237-018-0073-9

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models*. Cambridge, UK: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

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). Boca Raton, FL: CRC Press.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers No. No. 133). Paris, France: OECD Publishing.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*, 608–6262. https://doi.org/10.1037/a0034716

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

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173–183. https://doi.org/10.1080/08957347.2016.1171766

Guo, J., Gabry, J., & Goodrich, B. (2018). *Rjags: R interface to Stan. R package version 2.18.2*. Retrieved from https://CRAN.R-project.org/package=rstan

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1–17. https://doi.org/10.1111/j.2044-8317.2005.tb00312.x

Jakwerth, P. M., & Stancavage, F. B. (2003). *An investigation of why students do not respond to questions* (NAEP Validity Studies Working Paper No. NCES-WP-2003-12). Palo Alto, CA: American Institutes for Research.

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *American Statistician*, *52*(2), 93–100. https://doi.org/10.1080/00031305.1998.10480547

Köhler, C., Pohl, S., & Carstensen, C. H. (2015a). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling*, *57*(4), 499–522.

Köhler, C., Pohl, S., & Carstensen, C. H. (2015b). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement*, *75*(5), 850–874. https://doi.org/10.1177/0013164414561785

Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement*, *54*(4), 397–419. https://doi.org/10.1111/jedm.12154

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). *Omitted and not-reached items in mathematics. In 1990 National Assessment of Educational Progress (CRE Technical Report No. 347)*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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), 1–24. https://doi.org/10.1186/s40536-014-0008-1

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, *48*(3), 477–482. https://doi.org/10.1007/BF02293689

Lüdtke, O., Robitzsch, A., & Wagner, J. (2018). More stable estimation of the STARTS model: A Bayesian approach using Markov chain Monte Carlo techniques. *Psychological Methods*, *23*(3), 570–593. https://doi.org/10.1037/met0000155

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, *60*(4), 523–547. https://doi.org/10.1007/BF02294327

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521–538. https://doi.org/10.1177/0146621609355451

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195–215. https://doi.org/10.1007/BF02295283

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Report No. RR-96-30-ONR). Princeton, NJ: Educational Testing Service.

Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 205–228. https://doi.org/10.1111/bmsp.12117

Molenaar, D., & De Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika*, *83*(2), 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*(5), 606–626. https://doi.org/10.1080/00273171.2016.1192983

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197–219. https://doi.org/10.1111/bmsp.12042

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 113–162). Boca Raton, FL: CRC Press.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A*, *162*(2), 177–194. https://doi.org/10.1111/1467-985X.00129

OECD (2013). *Technical report of the survey of adult skills (PIAAC)*. Paris, France: Author. Retrieved from https://www.oecd.org/skills/piaac/_TechnicalReport_17OCT13.pdf

OECD (2017). *PISA 2015 technical report*. Paris, France: Author. Retrieved from https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg, Germany: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, *74*(3), 423–452. https://doi.org/10.1177/0013164413504926

Pohl, S., Ulitzsch, E., & von Davier, M. (2019). Using response time models to account for not-reached items. *Psychometrika*, *84*(3), 892–920. https://doi.org/10.1007/s11336-019-09669-2

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

R Development Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

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*(1), 74–104. https://doi.org/10.1080/15305058.2016.1231193

Robitzsch, A. (2014). On non-ignorable consequences of (partial) ignoring of missing item responses in large-scale assessment. In B. Suchan, C. Wallner-Paschon, & C. Schreiner (Eds.), *PIRLS & TIMSS 2011 – Competencies in reading, mathematics and science at the end of primary school: Austrian expert report* (pp. 55–64). Graz, AT: Leykam.

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* (Doctoral dissertation). Friedrich-Schiller-Universität Jena, Princeton, NJ. Retrieved from https://d-nb.info/1036873145/34

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (ETS Research Report No. RR-10-11). Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Sachse, K. A., Mahler, N., & Pohl, S. (2019). When nonresponse mechanisms change: Effects on trends and group comparisons in international large-scale assessments. *Educational and Psychological Measurement*, *79*(4), 699–726. https://doi.org/10.1177/0013164419829196

San Martín, E., del Pino, G., & de Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, *30*(3), 183–203. https://doi.org/10.1177/0146621605282773

Schnipke, D. L. (1996, April). *How contaminated by guessing are item-parameter estimates and what can be done about it?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

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

Stan Development Team (2017). *Stan modeling language: User's guide and reference manual (version 2.17.0)*. Retrieved from https://github.com/standev/stan/releases/download/v2.17.1/stan-reference-2.17.1.pdf

Ulitzsch, E., von Davier, M., & Pohl, S. (in press). *A multi-process item response model for not-reached items due to time limits and quitting*. Educational and Psychological Measurement.

Ulitzsch, E., von Davier, M., & Pohl, S. (2019). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2019.1643699

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Verbić, S., & Tomić, B. (2009). *Test item response time and the response likelihood*. Preprint, arXiv:0901.4356.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

Weeks, J. P., von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling*, *58*(4), 671–701.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61. https://doi.org/10.1111/emip.12165

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L., Kingsbury, G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, BC.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185–205. https://doi.org/10.1080/08957340902754650

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost (Ed.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). Münster, Germany: Waxmann.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Scaling PIAAC cognitive data. In OECD, *Technical report of the survey of adult skills (PIAAC)*. Retrieved from http://www.oecd.org/skills/piaac/Technical_Report_2nd_Edition_Chapters_17-23.pdf

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 646–661. https://doi.org/10.1080/10705511.2018.1545232

Received 29 January 2019; revised version received 21 June 2019

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1. Special cases of the SA+E model, Stan code, parameter recovery, and posterior predictive checks.