Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01.05
Volume 14 (2022), No. 1, 89–121

Gesa Brunn, Fritjof Freise, & Philipp Doebler

Modeling a Smooth Course of Learning and Testing Individual Deviations From a Global Course

Abstract

Formative assessment supplies valuable feedback for teachers and learners, and has been facilitated by computerized implementations. While longitudinal within-student assessment or within-class comparisons are useful, a normative interpretation of an individual's course of learning can only be given relative to a reference population. As current computerized assessment systems sample items from pools or adapt tests, monitored students might work on non-overlapping item sets, so that classic sum scores cannot be compared directly. To meet this challenge, the Smooth Growth and Linear Deviations Rasch Model (SGLDRM) is introduced, an extension of Rasch's item response theory model for binary test data. With the help of spline functions a smooth global course of learning is included. The model is flexible enough to accommodate increases and/or decreases of the mean ability level, which might be more or less pronounced at each measurement occasion. On the individual level, a random slope and a random intercept with amenable interpretations modify the global course of learning. Two measurement occasions suffice to estimate person-specific courses. A likelihood ratio test allows identifying students whose performance differs from the mean course. The methodology is illustrated with data from an online dyscalculia assessment and training.

Keywords

item response theory, latent growth curve model, formative assessment, random slope random intercept model, smooth growth curve

Gesa Brunn, ORCID: 0000-0002-0570-4520 · Prof. Dr. Philipp Doebler (corresponding author), ORCID: 0000-0002-2946-8526, Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany
email: gesa.brunn@tu-dortmund.de
doebler@statistik.tu-dortmund.de

Dr. Fritjof Freise, ORCID: 0000-0002-8493-8359, Institute for Biometry, Epidemiology and Information Processing, University of Veterinary Medicine Hannover, Bünteweg 2, 30559 Hannover, Germany
email: Fritjof.Freise@tiho-hannover.de

Modellierung eines glatten Lernverlaufs und Testung individueller Abweichungen von einem globalen Verlauf

Zusammenfassung

Formatives Assessment liefert Lernenden und Lehrenden wertvolles Feedback und ist durch computergestützte Implementationen stark vereinfacht worden. Zwar sind längsschnittliche individuelle Assessments und Vergleiche innerhalb einer Klasse nützlich, aber normative Interpretationen von individuellen Lernverläufen können nur relativ zu einer Referenzpopulation gegeben werden. Da aktuelle computergestützte Assessment-Systeme Items aus Pools zufällig auswählen oder Tests adaptieren, arbeiten die Getesteten u. U. auf sich nicht überlappenden Itemmengen, wodurch klassische Summenscores nicht direkt vergleichbar sind. Um dem zu begegnen, wird das Smooth Growth and Linear Deviations Rasch Model (SGLDRM) eingeführt, eine Erweiterung des Rasch-Modells für binäre Testdaten aus der Item-Response-Theorie. Durch Splines wird ein glatter globaler Verlauf eingebunden. Das Modell ist flexibel genug, um Anstiege und Verringerungen des mittleren Fähigkeitsniveaus abzubilden, welche je nach Messzeitpunkt unterschiedlich stark ausgeprägt sein dürfen. Auf der individuellen Ebene wird der globale Lernverlauf durch gut interpretierbare zufällige Achsenabschnitte und Steigungen modifiziert. Zwei Messzeitpunkte reichen aus, um personenspezifische Verläufe zu schätzen. Ein Likelihood-Quotienten-Test erlaubt es, Lernende zu identifizieren, die vom mittleren Lernverlauf abweichen. Die Methode wird anhand von Daten aus einem Online-System zur Diagnostik und Behandlung von Dyskalkulie illustriert.

Schlagworte

Item-Response-Theorie, latentes Wachstumskurvenmodell, formatives Assessment, Random-Slope-Random-Intercept-Modell, glatte Wachstumskurve

1. Introduction

The possibility to efficiently implement and administer test items with computerized test platforms facilitates the routine assessment of learning development (e.g., Klinkenberg et al., 2011; Kuhn et al., 2018; Mühling et al., 2017; Souvignier et al., 2014), and has progressed to a stage where system output informs school teachers rather than research scientists (e.g., Schurig et al., 2019). Next to processing test data to extract information on the overall development, it can also be of interest to track individual progress, give feedback on the performance, and initiate remedial measures if necessary, in the tradition of assessment for learning (Black & Wiliam, 1998). In order to utilize the assets of computer testing, we focus on the already highly developed item response theory (IRT) which offers strong frameworks for model extensions.

IRT models for longitudinal data are becoming increasingly important to track educational progress on a global and individual level. IRT provides several frameworks to model subject-based growth over multiple occasions, many of them being extensions of the Rasch (1960) model for binary item responses. A line of research starting with the seminal paper of Rost and Spada (1983) focuses on latent differences between occasions (Andersen, 1985; Embretson, 1991; Fischer, 1973, 1976, 1989). Initially, the approach was intended for two fixed occasions, but subsequent models were more general. Noteworthy refinements include the generalization of this approach to the two parameter logistic case by Embretson (1997), which was further extended (Andrade & Tavares, 2005).

However, there are still several issues that should be addressed and improved when modeling longitudinal data. Especially in computerized test and training systems, data collection often does not happen at fixed discrete equally spaced occasions, be it due to randomness or by design. Adequate models should allow for many if not all time values in an interval. Therefore, it is convenient to consider the time component to be continuous, and observations to be snap-shots of the current ability (e.g., Hecht et al., 2019). By using classic parametric models to describe the global progression of performance, for example, linear or quadratic, a certain shape of the global growth curve is presumed. This can severely restrict model fit and lead to biased representations.

In this study, we propose a non-parametric method to estimate a smooth global trend based on splines. This approach enables us to detect even small unexpected changes over time. Although the main focus of modeling longitudinal data in IRT has been the growth of a population or several subgroups, subject-specific performance can be of interest, for example when psychometric tests are used as formative evaluation tools parallel to educational interventions. By allowing a linear deviation from the average growth for each subject, we can track the individual starting level and gain given by a random intercept and slope, respectively. A likelihood ratio test allows identifying individual courses that significantly deviate from the average growth. As a consequence, feedback is possible and the initiation of interventions based on the intensity and direction of the deviation.

Saha (2016) presents a similar approach: The Bayesian dynamic item response model with semi-parametric and smooth ability growth (DIR-SMSG) uses B-spline functions to estimate ability growth in a dynamic IRT (Wang et al., 2013) framework. Saha’s (2016) model works with a discrete time component and assumes ability growth to be monotone. The present approach avoids monotonicity assumptions, incorporating growth, set-backs, and learning boosts. In contrast to the approach presented here, the DIR-SMSG uses spline functions to estimate each person’s ability separately. While a flexible model for individual growth results, the approach requires relatively dense data on the person level. To ensure accessible interpretation of both the population ability growth and person-specific deviations from the average growth the Smooth Growth Linear Deviation Rasch Model (SGLDRM) assumes person-specific linear deviations from the global trend. In this case, only two or more measurement occasions are required in order to estimate a person-specific ability growth which can be beneficial when only relatively few data are observed on the person level. In terms of complexity, the SGLDRM is between longitudinal IRT models with strong linearity assumptions on the logit scale and the DIR-SMSG.

The population's development is an explicit part of the SGLDRM, while it is necessary to treat individual subject effects as random effects to avoid over-parametrization (Baayen et al., 2008; Hecht et al., 2019). Realizations of these random effects are not estimated when fitting the model (but their variances and covariances are). However, the estimates (in the form of conditional modes/best linear unbiased predictors, BLUPs; Robinson, 1991) can be obtained subsequently.

The remainder of this paper is organized as follows: The advantages of item-level scaling in the context of formative assessment are stressed in the Section 2 and it is explained why non-linear growth curves safeguard against potential problems. Next, the SGLDRM, an extension of the Rasch model for longitudinal data, is presented: The model features a global smooth curve and individual deviations. The model can be seen as a generalized additive mixed model, leading to an estimation approach. The subsequent Section 4 develops a likelihood ratio test which helps to discern whether individual courses of learning deviate from the global course. The test is evaluated by simulation. An application to data from an online sample of primary school students with math difficulties or dyscalculia illustrates the usefulness of the method and is contained in Section 5. We close with some remarks on the limitation of the method and its relationship to existing methods.

2. Flexible Scaling on the Item Level With Non-Linear Growth Curves

Scaling on item level is a central characteristic of IRT and it produces some clear advantages compared to scaling on (sub-)test level. This has been stressed especially in the context of longitudinal applications (Reise & Haviland, 2005), though Jabrayilov et al. (2016) caution, that sufficient test length is still needed: An IRT model with acceptable fit for at least 20 items is recommended to reach acceptable misclassification rates of latent change. More generally, an ideal item set for measuring change provides high Fisher information for relevant parts of the ability continuum, so that person-specific reliability of the change scores is acceptable. When scaling on the level of a whole test, for example, with raw scores, the comparison of test results requires the tests to be equally difficult and the response patterns to be similar in the sense that subjects with the same true score make mistakes on items with similar difficulty. Creating such parallel tests for intensive longitudinal situations is complex, but not impossible (e.g., Fuchs et al., 1984; Strathmann & Klauer, 2012).

In contrast, scaling on item level allows calibrating large item pools, each test taker only working on a subset of items (Kolen & Brennan, 2004). These subsets can even be assembled randomly, or – for more effective and efficient testing – be drawn adaptively in computerized adaptive tests (CATs; van der Linden & Glas, 2000). Adaptive testing avoids ceiling and floor effects within the limits of the test design. With item-level scaling it is possible to model data from tests with time limits as it is not required for tests to be of the same length. Note, that this will work under the premise that the difficulty of an item is not affected by the number of items assigned in a test. However, item position effects do affect item parameters and include phenomena such as within-test learning and fatigue, and have been documented by Le (2007), Debeer and Janssen (2013), Debeer et al. (2014), Nagy et al. (2018), and Wu et al. (2019). A crucial problem is a correlation of item parameters and item position, say if a linear test design is used in estimation, but item order is varied subsequently. Hence, testing systems should either not vary item position or vary item position in the calibration system.

The validity of individual items can be examined when scaling on item level, which is beneficial for the process of item construction. These advantages of scaling on the item level require an overlap of the item subsets administered to each person, so that item parameters are properly linked (Kolen & Brennan, 2004). Randomly sampling items is a typical way to achieve this. The discussion highlights further caveats.

Next to flexibility with respect to the item subsets, the approach we introduce is flexible with respect to sampling occasions as well as with respect to the shape of the global course of learning. The use of spline functions (Wood, 2017) to model the global learning course allows flexible adjustments making it more convenient to fit real data compared to parametric alternatives. Smooth functions are especially useful to model short-term changes, for example, setbacks after school holidays or intervention effects.

3. Smooth Growth and Linear Deviation Rasch Model

We assume that $t = 0$ indexes the first measurement occasion and that there are occasions $t = 0, \ldots, T$, not necessarily evenly spaced. We do not assume that all persons are measured at the same occasions, reflecting a missing data situation that is common in longitudinal data. We index a pool of items by $i = 1, \ldots, I$ and assume that each person responds to a (maybe empty) subset of items at a measurement occasion and this subset might or might not be identical for all persons $j = 1, \ldots, J$.

The SGLDRM is introduced now. Let Y_{jit} be a dichotomous random variable that indicates whether at measurement time t person j answered item i either correctly ($Y_{jit} = 1$) or not ($Y_{jit} = 0$). Further assume that over the course of time a person's ability ϑ_{jt} can change while item difficulties β_i stay invariant. As the probability of solving an item should increase with a person's ability and decrease the more difficult an item is, the probability is assumed to increase monotonously with $\vartheta_{jt} - \beta_i$. The Rasch model assumes a logit-linear relationship:

$$P(Y_{jit} = y_{jit}) = \frac{\exp(y_{jit}(\vartheta_{jt} - \beta_i))}{1 + \exp(\vartheta_{jt} - \beta_i)}. \tag{1}$$

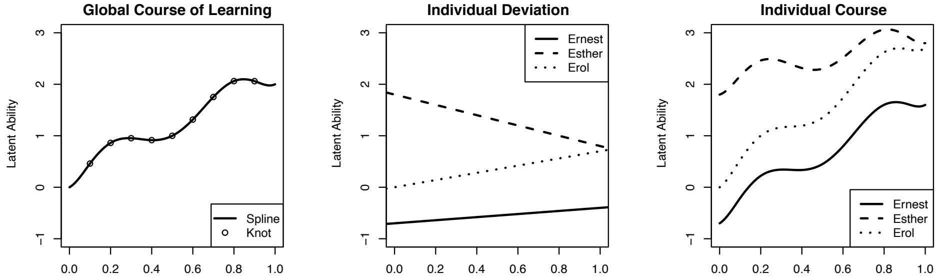
In our extension of the Rasch model, the latent person ability at time *t* is assumed to be composed of a latent global trend λ(*t*), in addition to a person-specific latent intercept δ_j and slope γ_j, such that

,

$$\vartheta_{jt} = \lambda(t) + \delta_j + \gamma_j t. \tag{2}$$

Note, that $\delta_j + \gamma_j t$ is a person-specific linear function in *t* that is interpreted as the deviation of person *j* from the global trend. We assume that the function of global mean latent ability on time, λ(*t*), is smooth, that is, we rule out sudden jumps. Otherwise, the function has an arbitrary shape which has to be estimated from the data and is not specified a priori. The technical approach via spline functions is detailed below. Figure 1 illustrates the approach.

Figure 1: Illustration of the Model: A Smooth Global Course is Modified by a Linear Individual Deviation (Synthetic Data for Three Students Named Ernest, Esther and Erol)



Similar to many other IRT models, local stochastic independence is assumed: Responses made by person *j* are independent conditional on person ability at time points t_1 and t_2 (where $t_1 = t_2$ is possible). Hence, conditional on item abilities at t_1 and t_2, the answer to item i_1 at t_1 does not affect the answer for item i_2 at t_2 (where $i_1 = i_2$ is possible), which can be expressed as

$$P(Y_{ji_1t_1} = y_{ji_1t_1} | \vartheta_{jt_1}, \vartheta_{jt_2}) = P(Y_{ji_1t_1} = y_{ji_1t_1} | Y_{ji_2t_2} = y_{ji_2t_2}, \vartheta_{jt_1}, \vartheta_{jt_2}). \tag{3}$$

Similarly, the probability of person j_1 solving item *i* correctly does not depend on whether person $j_2 \neq j_1$ was able to solve item *i*, leading to an independence assumption for persons.

If the same item is used at two different time points, the model, as specified, ignores potential local item dependence, that is, a person's chance to correctly answer at t_2 is conditionally independent on whether the item was solved at t_1 or not.

This might be most plausible if the item index represents a whole family of items (see the sample application), or for domains like mental calculation, attention, or clerical speed, while the assumption is implausible for items requiring some sort of insight or factual knowledge. Our recommendation is hence to avoid reusing the very same items if the SGLDRM is to be applied.

To complete the model specification of the SGLDRM, a bivariate normal marginal distribution for the latent variables δ and γ is assumed:

$$(\delta, \gamma)' \sim \mathrm{N}_2(\mathbf{0}, \Sigma), \tag{4}$$

where

$$\Sigma = \begin{pmatrix} \sigma_\delta^2 & \rho\sigma_\delta\sigma_\gamma \\ \rho\sigma_\delta\sigma_\gamma & \sigma_\gamma^2 \end{pmatrix} \tag{5}$$

is a not necessarily diagonal 2 × 2 covariance matrix. It is possible to constrain $\rho = 0$ for ease of fitting the model, which might however be implausible in some contexts. The individual latent variables $(\delta_j, \gamma_j)'$ are independent copies of the bivariate random variable $(\delta, \gamma)'$.

While the interpretation of ϑ_{jt} and β_i is familiar from the Rasch model, it is important to note how to interpret the other parts of the SGLDRM: The global latent growth curve λ(t) coincides with the latent ability of an average person with $\delta_j = \gamma_j = 0$. The parameter σ_δ^2 is the variance of latent ability at $t = 0$. The standard deviation σ_δ of latent ability is on the familiar logit scale and is hence easier to interpret. The larger σ_δ, the larger the spread of ability at $t = 0$. If $\gamma_j = 0$, the person-specific course of learning is just shifted up or down by δ_j. Hence, the larger σ_γ, which is the standard deviation of the person-specific slope of the linear deviation, the less likely it is for the individual trajectories to be practically parallel to the global curve given by $\lambda(t)$. Finally, ρ, the correlation of δ and γ, is to be interpreted depending on its sign: If ρ is positive, persons starting above the latent mean at $t = 0$ ($\delta_j > 0$) have larger slopes γ_j on average than those below the latent mean at $t = 0$ (“The rich get richer and the poor get poorer.”). If ρ is negative, persons starting above average ($\delta_j > 0$) tend to have smaller slopes than average ($\gamma_j < 0$), implying that, on average, persons with low ability catch up (to some extent) and that the latent variance might decrease as t increases.

3.1 The SGLDRM as a Generalized Additive Mixed Model

For purposes of parameter estimation, it is useful to understand the SGLDRM as a binomial generalized additive mixed model (binomial GAMM; Wood, 2017) with a logit link function, where $\lambda(t)$ is a smooth function of the covariate *time t*. The latent intercept δ and slope γ are random effects from the GAMM perspective and the item difficulties β_i are treated as fixed effects. Some details on binomial GAMMs follow, so that the connection becomes apparent.

A binomial GAMM with a univariate smooth function s is of the general form (Wood, 2017)

$$\mathrm{g}(\mathrm{E}(\mathrm{y}_\mathrm{l}|\mathbf{b})) = \mathbf{X}_\mathrm{l}\mathbf{a} + \mathrm{s}(\mathrm{x}_\mathrm{l}) + \mathbf{Z}_\mathrm{l}\mathbf{b}. \tag{6}$$

In the GAMM context, g is a differentiable link function with logit and probit being the most popular choices. The logit is chosen here, as this preserves the log-odds interpretation of many parameters familiar from Rasch models. On the right-hand side of the equation we have a model matrix of fixed effects $\mathbf{X}$ and the corresponding parameter vector $\mathbf{a}$. Here, only the item difficulties β_i are fixed effects, but the GAMM perspective allows for straightforward inclusion of other fixed person or item effects.

The smooth component $\mathrm{s}(\cdot)$ is included for flexible modeling of the effect of time on the learning process. Polynomials of high orders seem to offer a similar amount of flexibility, but overfitting polynomials leads to implausible oscillating patterns, limiting the ability to predict ability by interpolation. Instead, we use spline functions to approximate the process. They are highly flexible, smooth, and rather simple to use. Spline functions are reviewed with some detail in Subsection 3.2.

The random person effects δ_j and γ_j for the linear deviations in the SGLDRM are represented by the vector of random effects $\mathbf{b}$ and the corresponding model matrix of random effects $\mathbf{Z}$. By not treating them as fixed effects, the person-specific coefficients $(\delta_j, \gamma_j)'$, $j = 1, \ldots, J$, are not estimated when fitting the model. From a technical point of view, the person parameters are nuisance parameters and by treating them has random effects, they do not have to be estimated. However, after model fitting, individual estimates can be obtained, which we detail below. By eliminating person parameters from the model, the model fit does not depend on the particular set of people, but applies to subjects with abilities from the same distribution.

3.2 Regression Splines

Regression spline functions are a composition of continuous functions joined together at so-called knots to fit a smooth function to a certain set of noisy data (Wood, 2017). Instead of determining the shape of the mean course of learning in advance, for example, by choosing a linear function or another polynomial of some degree, the use of spline functions provides a more flexible fit to the data points and therefore is presumably closer to the "true" course of development. The true course is, of course, unknown since it is a latent psychological construct. The existence of a smooth underlying function of the true course is an assumption.

One issue in fitting a smooth function to data is to compromise between smoothness and fit. A class of splines that are commonly used as default in implementations are thin plate regression splines (Wood, 2003). They are based on the idea of minimizing the squared distance between spline and actual data, penalized by an additive term that gets larger the “wigglier” a spline gets, the so-called penalized residual sum of squares. A detailed illustration of spline functions can be found in Green and Silverman (1994).

When fitting a model with penalized regression splines, one has to choose the dimension of the basis that spans the space of spline functions. While a dimension which is too small can deteriorate a model’s fit, a higher dimensionality primarily leads to a higher computational cost of the thin plate regression spline (Wood, 2017). By choosing a basis with a sufficiently high dimension the fitted curve’s flexibility is unrestricted, while overfitting (too much “wiggliness”) is prevented by determining the penalty tuning parameter by a generalized cross-validation scheme. In other words, choosing a high dimension is not critical in packages with penalized splines, and this is one of the reasons we recommend the packages gamm4 (Wood & Scheipl, 2017) and mgcv (Wood, 2017) to fit the SGLDRM. Wood (2017, Chapter 5.9) provides an insight into the procedure of choosing and checking the basis dimension. A guideline given by Gu and Kim (2002) says that the basis dimension should be around $10n^{\frac{2}{9}}$ where n is the number of observations. Additional information on the estimation procedure and suggestions for implementations can be found in the Appendix.

As soon as the model is fitted to a sufficiently large set of data based on a representative set of subjects, the mean course of learning can, among other purposes, be used to compare the performances of newly tested individuals to it. This point is taken up in the next section.

4. Deviations From the Mean Course of Learning

After a baseline assessment, individual person parameters can be estimated, and point estimates together with asymptotic (e.g., Lord, 1983) or exact confidence intervals (Doebler et al., 2013; Klauer, 1991) can be used to assess initial ability. We now discuss a method for when longitudinal data is available: The aim is to compare results of an individual participant and its change over time to the mean development of the population. The SGLDRM composes a person’s ability at a certain point in time by a global trend and a person-specific intercept and slope. This composition can be used to develop statistical hypotheses regarding the divergence of a person’s ability from the global trend. Some approaches to hypothesis testing are presented below.

Note, that the inference in the following section is conditional given a fixed person j and, hence, the individual random effects δ_j and γ_j can be considered to be fixed parameters. Additionally, $\lambda(t)$ as well as item parameters β_i, $i = 1, \ldots, I$, are assumed to be known. Since we consider the deviation of an individual from the mean, all inference is done conditional on this person. The person index *j* is omitted in this section since a particular person's data is not necessarily part of the data set that is used to obtain the global course.

Assuming that the average course of learning $\lambda(t)$ represents typical or healthy development, it is important to check whether an individual's course of learning matches the mean course of learning relating to a certain population. If this was the case this person parameter, denoted by ϑ_t, would only consist of the global trend for all points in time *t*:

$$\vartheta_t = \lambda(t). \tag{7}$$

With respect to the person-specific linear deviation this yields the hypothesis

$$H_0 : \delta = \gamma = 0. \tag{8}$$

If this null hypothesis is discarded the individual either would have a level of ability above or below average ($\delta_j \neq 0$), or learn faster or slower than the average ($\gamma_j \neq 0$) or a combination of both. Figure 2 shows some theoretical learning courses.

To test the hypothesis in (8), we propose to employ a likelihood ratio test (LRT), which compares the likelihood functions under H_0 and the alternative H_1. We first make the likelihood $L(\delta,\gamma \mid \mathbf{y})$ explicit for a pair $(\delta,\gamma)'$ given the data $\mathbf{y}$ of person *j*: Assume test performance of person *j* has been observed at times t_k, $k = 1, \ldots, K$. The items administered to the individual at hand might not be the same at each point in time, so denote the index sets containing the item indices by A_{t_k}, $k = 1, \ldots, K$. Thus, one observes $y_{it_k} \in \{0,1\}$ for $i \in A_{t_k}$ and $k = 1, \ldots, K$. Local independence assumptions then imply

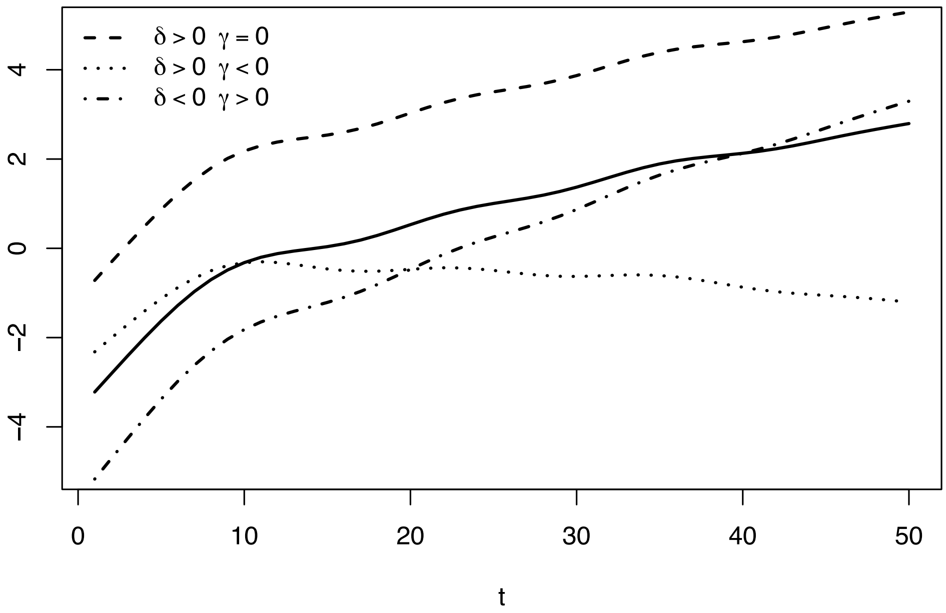
$$L(\delta,\gamma|\mathbf{y}) = \prod_{k=1}^{K} \prod_{i \in A_{t_k}} \frac{\exp(y_{it_k}(\vartheta_{t_k} - \beta_i))}{1 + \exp(\vartheta_{t_k} - \beta_i)}. \tag{9}$$

The ratio of the likelihoods is then given by

$$r(\mathbf{y}) = \frac{L(0,0|\mathbf{y})}{L(\hat{\delta},\hat{\gamma}|\mathbf{y})} \tag{10}$$

where numerator and denominator are the maximum of the likelihood under H_0 and the alternative, respectively. The value of the latter is the likelihood taken at the ML estimate, $(\hat{\delta},\hat{\gamma})'$, which is the conditional mode/BLUP.

Figure 2: Illustration of a Smooth Global Trend (Solid Line) and Three Exemplary Deviations (Broken Lines)



The smaller the ratio $r(\mathbf{y})$, the less likely it is that H_0 is true given the observations $\mathbf{y}$. Under H_0 the statistic $W = -2\ \ln(r(\mathbf{y}))$ is asymptotically χ_2^2-distributed (Casella & Berger, 2002). In finite samples, especially when few items are used, the χ^2-approximation can fail. The finite sample performance can be improved by an empirical Bartlett correction: Rather than comparing the test statistic $W = -2\ \log(r(\mathbf{y}))$ to the corresponding quantile of the χ_2^2-distribution, it is first multiplied by an estimate of the factor $2/\ \mathrm{E}(W)$, so that the corrected test statistic estimates (Pawitan, 2001)

$$\mathrm{W}^* = \frac{2\mathrm{W}}{\mathrm{E}(\mathrm{W})}. \tag{11}$$

4.1 Variants of the LRT

There are two potentially useful variants of the LRT: Assume an educator is interested to find out whether (a) a child manages to develop parallel to the mean course (i.e., is the distance to the global mean course stable?), or (b) whether a child under- or overperforms given what growth is expected by the child's initial ability. Mathematically, the first variant (a) is straightforward: The course is parallel if, and only if, $\gamma = 0$. So maximizing $L(\hat{\delta}\,\mathit{parallel},\ 0 \mid \mathbf{y})$ as a function of $\hat{\delta}\,\mathit{parallel}$ and comparing to the ML-estimate yields the variant of the test statistic. Setting

$$\mathrm{r}_{\mathrm{parallel}}(\mathbf{y}) = \frac{\mathrm{L}(\hat{\hat{\delta}}_{\mathrm{parallel}}, 0|\mathbf{y})}{\mathrm{L}(\hat{\delta}, \hat{\gamma}|\mathbf{y})} \tag{12}$$

we define $W_{parallel} = -2 \ln(r_{parallel}(\mathbf{y}))$. Then $W_{parallel}$ is approximately χ^2_1-distributed under H_0: $\gamma = 0$ and one can proceed as before. We mention in passing that a reference value of $\gamma \neq 0$ could also be tested.

For the second variant (b) first note that when the bivariate normal distribution assumption in the SGLDRM holds (with $E[\delta] = E[\gamma] = 0$), the conditional expectation of γ given δ is given by

$$\mathrm{E}(\gamma|\delta) = \rho\frac{\sigma_\gamma}{\sigma_\delta}\delta \tag{13}$$

by standard results for the bivariate normal distribution. The parameters in this expression are taken from (estimates) of the covariance matrix in Equation 5. The conditional expectation reflects the best guess of γ given δ, and hence reflects the growth one would expect given the initial performance. The second variant will hence be called the conditional variant, and we set

$$\mathrm{r}_{\mathrm{cond}}(\mathbf{y}) = \frac{\mathrm{L}(\hat{\hat{\delta}}_{\mathrm{cond}}, \mathrm{E}(\gamma|\hat{\hat{\delta}}_{\mathrm{cond}})|\mathbf{y})}{\mathrm{L}(\hat{\delta}, \hat{\gamma}|\mathbf{y})}, \tag{14}$$

and define $W_{cond} = -2 \ln(r_{cond}(\mathbf{y}))$. As in the parallel case, W_{cond} is approximately χ^2_1-distributed under H_0: $\gamma = E(\gamma \mid \delta)$ and one can proceed as before. However, the null hypotheses for $W_{parallel}$ and W_{cond} now contain more than one point. This means no empirical Bartlett correction can be implemented, since neither $E(W_{parallel} \mid H_0)$ nor $E(W_{cond} \mid H_0)$ are well defined.

4.2 Power of the Person-Level LRT and its Variants

In a simulation, we want to examine the performance of the proposed LRTs for some cases of individual deviation as well as the behavior under the null hypotheses. We focus on the LRT for H_0: $\delta = \gamma = 0$, but provide some insight into the variants.

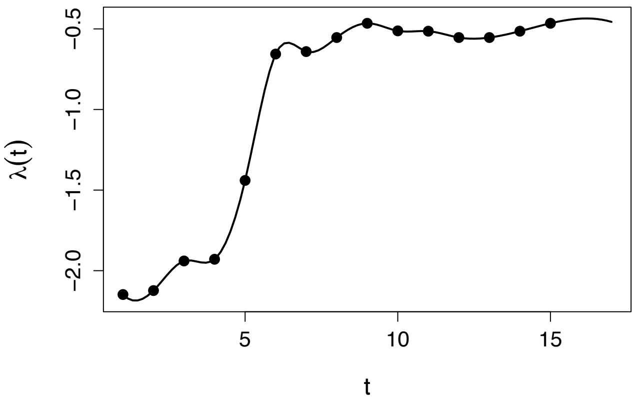
4.2.1 Power of the LRT and Calibration Error

In a first step, assuming a nonlinear mean course of learning and certain values for Σ, the covariance matrix of the individual effects, a calibration sample is generated. To explore the frequency of rejection of the true null hypothesis (type I error), the calibration data set is generated from Equations 1, 2, 4, and 5 with J subjects, $I = 5$ items, and $T = 10$ measurements per subject. The individual intercept and slope are set to be uncorrelated (thus $\rho = 0$). The different settings for Σ are given in Table 1. To additionally examine the LRT properties without a calibration error, comparisons made in setting N are based on the true parameters rather than estimations made on the calibration data. Since we would like to use the spline functions to their full capacity, we refrain from choosing a polynomial function as global learning course $\lambda(t)$. Instead, we arbitrarily choose 15 numbers and arrange them in ascending and descending sequences. We interpolate this pseudo data with a spline function and use this spline function as $\lambda(t)$. The data points as well as the spline are illustrated in Figure 3.

Table 1: Values of σ_δ and σ_γ for Simulated Data With $\rho = 0$ and Mean Vector **0** (see Equations 4 and 5). Values for β_i, $i = 1, \ldots, 5$ are set to $-1, -0.5, 0, 0.5, 1$ for all Calibration Sets (Including Setting N)

Calibration set	σ_γ	σ_δ	J
A	0.2	1	100
B	0.2	1	2500
C	0.75	0	2500
No calibration error (N)	–	–	–

Figure 3: Arbitrary Global Course Used for the Generation of Simulation Data

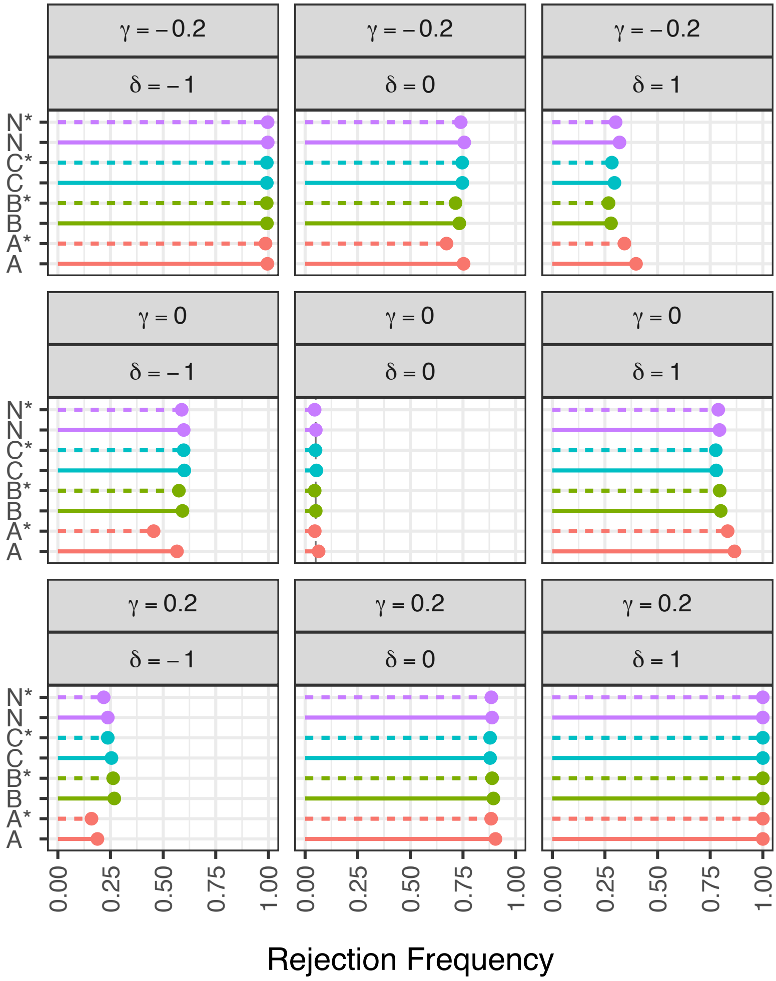


Note. The numbers used for interpolation are −2.15, −2.12, −1.94, −1.93, −1.44, −0.66, −0.64, −0.55, −0.47, −0.51, −0.51, −0.55, −0.55, −0.51, −0.47.

After fitting the model (Equation 6) to the simulated calibration data set an estimate (t) of the smooth course of learning results. The fitted global trends are shown in Figure 5. With only 100 virtual individuals from data set A the overall shape of the development is captured albeit smoother and with less precision than with higher sample sizes (mean bias in the fixed effects: A: −0.122, B: −0.005, C: −0.006). An increasing variability in individual slopes seems to have no substantial impact on the estimation of λ.

Data points for new individuals are simulated by drawing correlated Bernoulli responses with conditional probabilities shown in Equation 1 for fixed values of γ and δ. To examine hypothesis test behavior under H_0, a first data set is drawn with γ = δ = 0. The LRT statistic, as presented in the previous section, is calculated by using the estimated mean learning curve $\hat{\lambda}(t)$ of the calibration data set and individual ML estimates $\hat{\delta}_j$ and $\hat{\gamma}_j$. This procedure is repeated 5000 times. Figure 4 shows the percentage of test statistics which are greater than the 95% quantile of

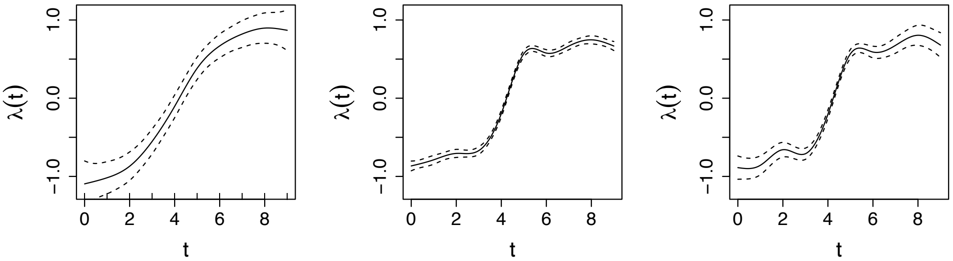
Figure 4: Rejection Frequencies of LRT Test at Significance Level α = 5% for Data Sets A, B, and C and Different Types of Deviation



Note. Bartlett-corrected performances are asterisked.

the χ^2-distribution. Under the null hypothesis (see center panel), it shows a rejection frequency that is slightly higher than 5% for all of the three calibration data sets (6.72%, 5.16%, 5.36%, respectively) as well as the comparison to the true parameters (5.10%). The significance level is exceeded. The Bartlett correction improves the results slightly for all data sets (4.52%, 4.56%, 5.18%, respectively) and setting *N* (4.54%). Table 2 holds bias, root mean squared error (RMSE), and the coverage of parameter estimations for the four simulation settings.

Figure 5: Average Courses λ(t) Fitted for Data Sets A, B, and C, Respectively



To examine the behavior under H_1, the values for δ and γ are varied systematically: The parameters take three different values respectively, one negative (δ = −1, γ = −0.2), one neutral (δ = γ = 0), and one positive (δ = 1, γ = 0.2). The test seems to find deviations more consistently if the deviations of both parameters are directed the same way. For example, a deviation of a student whose course of learning is both shifted upwards and increases more rapidly than the global course, for example, δ = 1, γ = 0.2, is very likely to be detected. However, individuals with an above-average intercept and a below-average slope are detected only about 30% of the times.

The biases in δ and γ are large for some cases, most pronounced when δ = −1 and γ = −0.2. The deeper reason are floor effects, that is, the test material is way too hard for the simulated persons. One can see the combination of δ = −1 and γ = −0.2 as a misfit of the item set and the person. As seen in Figure 3, the mean trajectories start in the range of −1.2 to −0.9. A person with δ = −1 hence starts at an ability level of −2.2 to −1.9 and hence the chance to solve an item of average difficulty is smaller than 13%. Over the course of the ten simulated measurement occasions, the mean gain is up to 2 units, depending on the mean trajectory in Figure 3. Hence, at Time 10 a γ of −0.2 leads to a latent ability which cancels out the mean gain or is even a net ability decrease over time. As a consequence, the whole trajectory has to be estimated based on potentially many zero scores, so the maximization of the person likelihood in Equation 9 can yield extreme estimates of δ and γ. In other words, the substantial growth we simulate uncovers deficits of the procedure in pathological cases.

Table 2: Rejection Rates of LRTs, Bias, Root Mean Squared Error and Coverage of Asymptotic 95% Confidence Intervals of Parameter Estimates Under Different Settings

Setting/ data set	δ	γ	Rej. rate	Rej. rate Bartlett	Bias δ	Bias γ	RMSE δ	RMSE γ	Coverage δ	Coverage γ
N	−1.0	−0.2	0.999	0.998	−1.47	−5.79	18.81	17.46	0.96	1.00
	0.0	−0.2	0.865	0.854	−0.46	−0.10	5.91	2.51	0.97	0.98
	1.0	−0.2	0.315	0.297	−0.06	0.00	0.87	1.21	0.96	0.96
	−1.0	0.0	0.598	0.588	−0.76	0.07	7.19	1.30	0.96	0.97
	0.0	0.0	0.051	0.045	−0.18	0.02	1.05	0.15	0.96	0.96
	1.0	0.0	0.794	0.788	−0.08	0.01	0.76	0.99	0.96	0.95
	−1.0	0.2	0.309	0.290	−0.26	0.03	1.24	1.24	0.96	0.96
	0.0	0.2	0.956	0.954	−0.18	0.03	0.93	0.27	0.96	0.96
	1.0	0.2	1.000	1.000	−0.11	0.03	0.78	0.78	0.97	0.97
A	−1.0	−0.2	0.997	0.994	−3.00	−5.81	29.22	17.86	0.93	0.99
	0.0	−0.2	0.834	0.778	−0.23	−0.13	8.08	2.63	0.93	0.97
	1.0	−0.2	0.376	0.315	0.25	−0.03	0.94	1.24	0.93	0.95
	−1.0	0.0	0.540	0.438	−0.43	−0.01	7.72	1.99	0.93	0.95
	0.0	0.0	0.067	0.045	0.14	−0.01	1.07	0.15	0.93	0.95
	1.0	0.0	0.871	0.839	0.25	−0.02	0.80	1.03	0.92	0.94
	−1.0	0.2	0.313	0.263	−0.00	0.01	1.27	1.22	0.94	0.94
	0.0	0.2	0.971	0.959	0.13	-0.00	0.92	0.24	0.93	0.95
	1.0	0.2	1.000	1.000	0.21	-0.00	0.83	0.82	0.93	0.95
B	−1.0	−0.2	0.997	0.997	−2.68	−5.41	24.15	17.20	0.96	0.99
	0.0	−0.2	0.846	0.833	−0.40	−0.20	4.73	3.42	0.97	0.98
	1.0	−0.2	0.281	0.268	−0.14	0.01	0.87	1.19	0.97	0.96
	−1.0	0.0	0.592	0.575	−0.61	0.03	3.62	1.73	0.97	0.97
	0.0	0.0	0.052	0.046	−0.22	0.03	1.02	0.15	0.97	0.97
	1.0	0.0	0.801	0.795	−0.12	0.02	0.77	0.99	0.96	0.95
	−1.0	0.2	0.340	0.334	−0.32	0.05	1.24	1.26	0.97	0.97
	0.0	0.2	0.962	0.959	−0.19	0.03	0.92	0.28	0.97	0.96
	1.0	0.2	1.000	1.000	−0.18	0.04	0.84	0.77	0.96	0.96
C	−1.0	−0.2	0.998	0.998	−2.60	−5.82	25.43	17.81	0.96	1.00
	0.0	−0.2	0.871	0.869	−0.82	−0.15	10.31	3.41	0.97	0.98
	1.0	−0.2	0.304	0.292	−0.13	0.01	0.88	1.20	0.96	0.96
	−1.0	0.0	0.628	0.627	−0.67	0.01	6.01	2.07	0.96	0.97
	0.0	0.0	0.054	0.052	−0.19	0.02	1.01	0.14	0.97	0.96
	1.0	0.0	0.764	0.763	−0.14	0.02	0.77	0.99	0.96	0.96
	−1.0	0.2	0.302	0.290	−0.29	0.03	1.29	1.25	0.96	0.96
	0.0	0.2	0.949	0.948	−0.21	0.03	0.93	0.27	0.96	0.97
	1.0	0.2	1.000	1.000	−0.17	0.04	0.84	0.78	0.96	0.96

Note. Rej. = rejection; RMSE = root mean squared error.

4.2.2 Behavior of Variants of the LRT

We compare the LRT with its variants in the setting with no calibration error (*N*). Recall, that the global growth curve covers a range of 2.5 logits and hence reflects extreme growth. We hence also employ the same curve multiplied with a factor of 0.25, reflecting moderate global growth, and use the same parameter settings and procedure otherwise. For W_{cond}, we use $\rho = 0.2$, $\sigma_\delta = 1$, and $\sigma_\gamma = 0.2$ to calculate $E(\gamma \mid \delta)$. For all tests the scenario $\delta = \gamma = 0$ is contained in the null hypothesis, and for $W_{parallel}$, all scenarios with $\gamma = 0$ are consistent with H_0. We determined the proportion of rejected null hypotheses for *W*, $W_{parallel}$, and W_{cond} in 5000 replications, and the results are presented in Table 3. Since the Bartlett correction is only applicable to *W*, we omit it in this simulation.

Table 3: Proportion of Rejected LRTs

True parameter			LRT (variant)		
δ	γ	λ	W	$W_{parallel}$	W_{cond}
−1	−0.2	1.00	.994	.996	.999
0	−0.2	1.00	.757	.720	.795
1	−0.2	1.00	.305	**.038**	**.055**
−1	0.0	1.00	.617	**.694**	.702
0	0.0	1.00	.052	.035	.055
1	0.0	1.00	.793	**.825**	.872
−1	0.2	1.00	.237	**.031**	**.063**
0	0.2	1.00	.882	.819	.888
1	0.2	1.00	1.000	1.000	1.000
−1	−0.2	0.25	.999	1.000	1.000
0	−0.2	0.25	.847	.826	.829
1	−0.2	0.25	.391	**.053**	**.056**
−1	0.0	0.25	.791	**.908**	.875
0	0.0	0.25	.047	.048	.054
1	0.0	0.25	.852	**.883**	.910
−1	0.2	0.25	.367	**.048**	**.043**
0	0.2	0.25	.878	.778	.859
1	0.2	0.25	1.000	1.000	1.000

Note. 5000 replications. Cases with excessive rejections of H_0 or low power in boldface.

The upper half of the fourth column of Table 3 is for the *W* LRT (both, γ and δ, are freely estimated under H_1). This reproduces a portion of Table 2, with minor discrepancies in the third decimal place due to Monte Carlo errors, so we do not repeat the above interpretation. Generally speaking, the variants of the LRT are able to detect violations of the null hypothesis, but there are scenarios which have low power (especially when the global course and the individual course intersect, i.e., $\delta = 1, \gamma = -0.2$ or $\delta = -1, \gamma = 0.2$). In some scenarios, compatible with the null hypothesis, $W_{parallel}$ will create false positive test results ($\delta = 1$ or $\delta = -1$). When the growth curve is extreme, the recovery of the person parameters is biased, affecting in turn the LRT variants. The effect is less pronounced when the growth is less extreme, but still observable (with 32% and 18% rejections when $\gamma = 0$ and $\delta = 1$ or $\delta = -1$, respectively). We recommend using the variants only when person parameter estimates can be assumed to be unbiased, severely limiting their applicability.

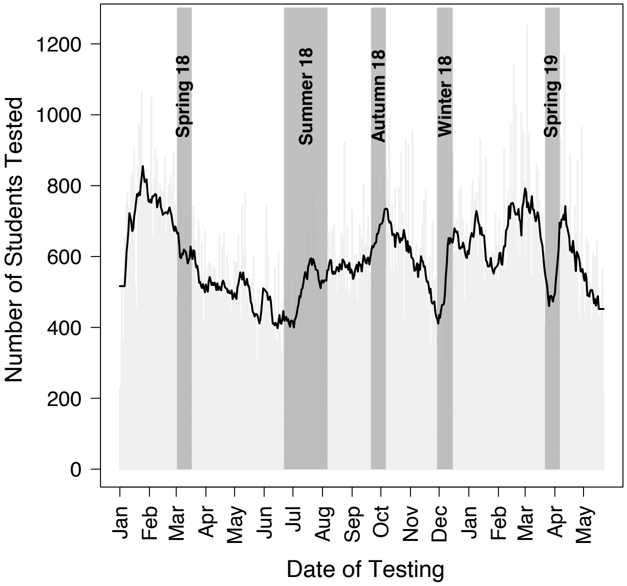
5. Empirical Illustration: Dyscalculia

Data from *Meister Cody von Talasia* (CODY) is reanalyzed. CODY is an online test and training system that was developed to monitor the learning progress of children with mathematical learning difficulties. More specifically, data from a numeracy test previously studied by Schwenk et al. (2017) is reanalyzed. The response accuracy is the target criterion; the two alternatives, response speed and response efficiency (cf. Schwenk et al., 2017), are not considered here. We stress that we are not measuring numeracy in an all-encompassing sense, since response times are not considered here.

Basic arithmetic skills are assessed with addition and subtraction items in a number range between 0 and 20. For each type of calculation task the test taker has 90 seconds to answer as many items correctly as possible. The items are randomly drawn from a pool of $N_{add} = 127$ and $N_{sub} = 143$ calculations and are presented to the user one by one. The tests go along with corresponding training sessions and are integrated into a motivational story. The data on hand was collected between January 2018 and June 2019 and comprises almost 300 000 individual observations made in about 14 500 tests on 3500 children. The usage frequency over the years and potential setbacks due to vacations are pictured in Figure 6. In order to make the SGLDRM applicable to the data, we transformed the time stamps and clustered the items. One should note that there is a variety of possibilities to transform and create the necessary variables and the following approach is taken for demonstration purposes.

Table 4 gives an idea of the longitudinal data set. Two time variables are part of the data set: a *day* variable giving the number of training sessions since the child's first registration and the *date* of the assessment. We combine these two variables to create a quasi-continuous time variable. For each student $t_j = 0$ indicates the time of their registration. We use *day* as a basis and add the time of day as a percentage. Students who are tested on their fifth day at 6 p.m. thus get a *time* value of $t_j = 4.75$.

Figure 6: Frequency of Application Usage Smoothed by Central Moving Average With Window Length 9 Days



Note. Vacation periods for North Rhine-Westphalia are added to indicate nationwide vacation periods.

Table 4: General Structure of Dyscalculia Data

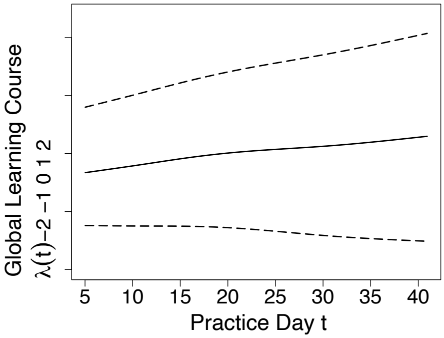
Index	Subject	Day	Calculation	Date	Task type	Correct
1	1	5	4 + 8 = 12	2018-05-13 10:03:08	add	0
2	1	5	9 + 9 = 18	2018-05-13 10:03:19	add	0
3	1	5	11 − 1 = 10	2018-05-13 10:03:30	sub	1
4	2	5	2 + 3 = 5	2018-02-14 15:27:57	add	1
5	2	5	1 + 1 = 2	2018-02-14 15:28:06	add	1
6	2	5	19 − 7 = 12	2018-02-14 15:28:13	sub	0
7	1	10	4 + 8 = 12	2018-05-20 16:05:50	add	1
8	1	10	9 + 9 = 18	2018-05-20 16:06:08	add	0

Note. add = addition; sub = subtraction.

There are 270 different calculation tasks in the item pool which would make their estimation computationally intense and consume much of the information that we would rather use to compute the overall course of learning. Instead we categorize the addition and subtraction items into three different groups, respectively: calculations within 10 (*add* ↓ 10, *sub* ↓ 10), between 10 and 20 (*add* ↑ 10, *sub* ↑ 10), and those that require passing the 10 (*add* ↕ 10, *sub* ↕ 10). In that way, there are only six item parameters to be estimated while we are still able to assess the difficulties of the item clusters.

For model fitting we only use observations from children tested on days 5, 10, 15, ..., as this time lag is recommended by the CODY authors, and who participated in at least six assessments. Apart from that, tests with "perfect" scores are excluded to maintain estimability and interpretation. Thus, we consciously use only 37% of the total observations for parameter estimation, so that the mean curve represents an average child requiring training and adhering to CODY training for at least 30 training days. Figure 7 shows the estimated mean curve that corresponds to $\lambda(t)$ of the SGLDRM. To have some kind of guideline for individual performance, the first and third quantile of the sample parameters are added to the curve. Table 5 shows the estimated item parameters. As anticipated, estimates suggest that addition and subtraction tasks in the number range from 0 to 10 are the easiest. Addition tasks seem to be easier than the according subtraction tasks.

Figure 7: Estimated Global Course of Learning for the Dyscalculia Example



Note. Dashed lines at the first and third sample quartiles of subject parameters δ and γ.

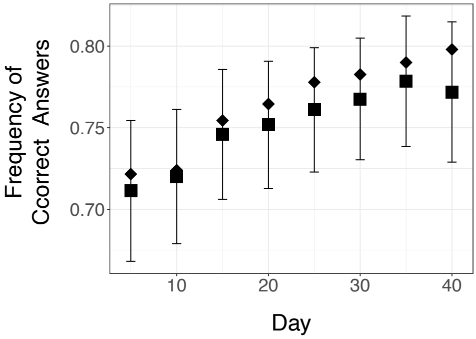
Table 5: Estimations of Item Parameters for the Dyscalculia Example for the SGLDRM

Item parameter	$\beta_{add\updownarrow 10}$	$\beta_{add\uparrow 10}$	$\beta_{add\downarrow 10}$	$\beta_{sub\updownarrow 10}$	$\beta_{sub\uparrow 10}$	$\beta_{sub\downarrow 10}$
estimation	1.25	1.53	2.69	0.61	0.99	2.12

Note. Increasing values indicate decreasing difficulty.

Additionally, we examine overall learning development by simply determining the averages of correct responses per day (see Figure 8). The development is similar to the curve that was estimated for the SGLDRM. Apart from a more defined increase the major difference is a decrease in performance from day 30 to day 40 of the mean when perfect scores are omitted (■), thereby excluding an increasing amount of correct responses as can be seen by comparing the development based on all observations (♦) and those without perfect scores (■). The non-parametric method gives each calculation the same weight in the estimation process, which might lead to discrepancies between the methods especially when only few items are responded to in each test run.

Figure 8: *Mean of Correctly Answered Items per Day of all Given Observations* (♦) *and Perfect Scores Omitted* (■)



Note. Bars marking ± one standard error of the mean with perfect scores omitted.

We give examples of individual analyses now. Three students are chosen in order to demonstrate different types of developmental patterns. We randomly choose one student each with the characteristics (a) average learner starting at average level, (b) extraordinary learner, and (c) extraordinary initial level. Those three pseudo groups are certainly no educated classifications but are provided for demonstration. Student A seems to have an average performance (see Figure 9) with δ_A = 0.534 and γ_A = −0.017. Their likelihood ratio test statistic (without Bartlett correction) of W_A = 1.095 is below the quantile $\chi^2_{2;0.95}$ = 5.991, thus the Hypothesis H_0: $\delta_A = \gamma_A = 0$ cannot be rejected with given significance level α = 5%. The non-parametric analysis (see Figure 10) gives the impression of an unsteady performance by Student A oscillating around the average. Both methods show Student B starting at an average level of performance which soon drops rapidly (δ_B = 0.754, γ_B = −0.085). The likelihood ratio test indicates this individual's learning curve to significantly differ from the average course $\lambda(t)$ with W_B = 18.020. The third subject, Student C, also shows a significant finding with a test statistic of W_C = 43.865. However, in contrast to Student B it is due to their particularly good initial level of performance with estimated parameters δ_C = 1.095 and γ_C = 0.000.

Figure 9: Individual Learning Courses of Student A (red), Student B (green), and Student C (blue) as Estimated in the SGLDRM Framework

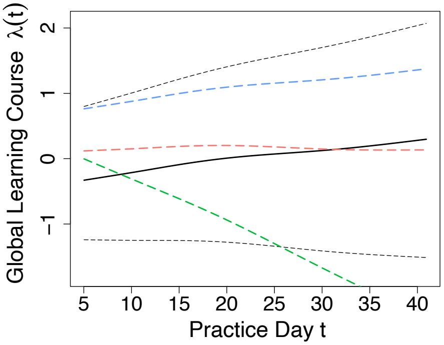
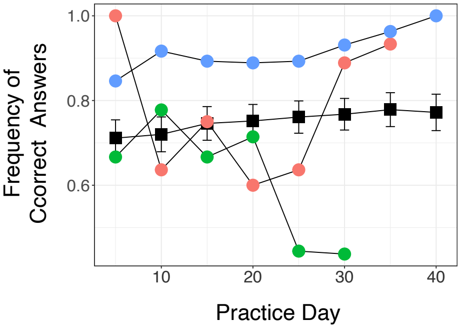


Figure 10: Individual Scores of Student A (red), Student B (green), and Student C (blue) in Comparison to Population Average Scores (black)



It is noteworthy that given analyses do not take any speed component into account. The number of completed items per test run are not used to estimate the students' abilities. This might be a great loss of information as can be seen in this example. While Student C works on 27.83 items on average with the overall mean of given data being 20.16 (*SD* = 5.65) their two fellow students do not do as well with 13.43 (Student A) and 9.33 (Student B) items per test. Although the performance of Student A might be average in regard to calculating accuracy, there might be a lack of speed in the performance. One naive method to deal with this inconvenience is to treat each item of the item pool that was not worked on in a test run as an incorrect answer. In that way it is equally valued to give an incorrect response or to not have the time to work on an item at all. However, this leads to different kinds of issues especially when working with great amounts of items and comparatively low sample sizes. To begin with, the computational costs rise up drastically, as the number of observations increases. Furthermore, estimated item parameters are not interpretable in practice as it is not clear if an incorrect response to an item was made due to high difficulty or if the item simply was not drawn into the test sample.

6. Discussion

The SGLDRM is a model for longitudinal data on the item level. We have demonstrated that a smooth global course of learning (CoL) can be estimated, which can often be interpreted as a reference CoL. In the SGLDRM, deviations from the global CoL are linear, so that the resulting person-level intercept and slope parameters can be estimated from two or more repeated observations. Potential applications include situations with a moderate number of repeated measurements per person. The measurement occasions can be unequally spaced. Also, items might be part of a larger pool of items. Regardless of which items are used, person-level deviations can be tested, allowing to detect deviations from the global course.

6.1 Relationship to Existing Approaches

Andersen's (1985) model expands the Rasch (1960) model by a time-dependent person parameter ϑ_{jt} for person j at measurement occasion t. Similarly, Embretson (1991) assumes a baseline person ability at the first measurement occasion to which the gain in ability is added at each occasion, so that the ability of person j at time t can be written as $\sum_{\mathrm{m}=1}^{\mathrm{t}} \vartheta_{\mathrm{jm}}$. In the linear logistic test model (LLTM; Fischer, 1973) change is modeled in regard to items instead of subjects. It is assumed that the difficulty of an item is composed of different basic (cognitive) operations that are needed in order to solve an item. Change over measurement occasions is analyzed by parameterizing items with a component that indicates learning effects, that is the item difficulty changes due to experiences that were made on items before. The linear logistic model with relaxed assumptions (LLRA; Fischer, 1976) additionally allows for item-specific person abilities.

In a more recent approach, Hecht et al. (2019) use a generalized linear mixed model (GLMM) framework for their continuous-time Rasch model for dichotomous responses similar to the present approach. GLMMs are a special case of GAMMs with the smooth function being a linear function. The authors treat item effects as fixed and person effects as random, however the relation and development of the person ability is motivated and implemented differently. The person ability of person j at (discrete) time t is modeled as an autoregressive process of order 1 (AR[1]), with an underlying continuous time model. This underlying model accounts for global intercept and slope which can be translated to our model by setting the global trend $\lambda(t)$ as a degree 1 polynomial. Hecht et al. (2019) consider a person-specific intercept, however no person-specific slope. This can be interpreted as individual developments running parallel to the average development.

The bivariate normal random effect in the SGLDRM implies that the variance of latent ability at time t is given by an expression quadratic in t. This implies that a norm based purely on the SGLDRM might misestimate latent variance, even if the spline correctly estimates the latent mean curve. If norms are required and sample size is sufficient, semiparametric continuous norming provides accurate recovery of latent variance in simulations (Lenhard et al., 2019).

In their generalized explanatory longitudinal item response model Cho et al. (2013) include the possibility to subdivide items and subjects into groups and estimate the corresponding averages for discrete time points. The authors use a GLMM framework with a logit link function to fit the model. With only one item group and one subject group the model consists of three subject-based and two item-based components. Person ability is composed of an overall intercept, an average ability for each time point, and a person-specific ability per time point in which the latter is modeled as a random effect. The item component consists of a fixed average and item-specific deviations from this average which are again taken as random effects. Similar to Fischer's models, the assumption of unidimensional items is relaxed. In order to solve an item, multiple cognitive skills are assumed to be required, differing among the items or rather item groups. This approach can be added to the SGLDRM. Fixed item group effects were technically implemented in the context of the dyscalculia example with six item clusters. Additional random item effects would have accounted for item-specific deviations.

6.2 Recommendations, Limitations and Potential Extensions

6.2.1 Item Overlap

One limitation of the fitting process is that item sets administered at different occasions need to overlap. Hence, designs with disjoint tests for each measurement occasion are not suitable for the SGLDRM, at least not without additional (equality) constraints on item parameters. Random sampling of items from a larger pool typically provides enough overlap. One benefit of this model is the possibility to compare student abilities and CoLs even when students worked on disjoint sets of items. An assumption of the SGLDRM is that item difficulty is invariant over time. Situations where difficulty drifts, say due to an intervention or schooling as usual, are not contained in the current approach. We recommend testing the item parameter invariance (e.g., Millsap, 2010).

Note, that the conditional independence assumption means that beyond current ability, no other factors govern test behavior. Especially, repeated measurement with the same items has to proceed with ample time between measurement occasions so that memory effects do not violate conditional independence. Items need to be drawn without replacement at the same measurement occasion to rule out local item dependence.

6.2.2 Time and Other Influences

The SGLDRM can index measurement occasions with fine granularity, that is, time can be continuous in the model. In the sample application to CODY data, a time of zero indicated the beginning of a training, regardless of actual date. As a consequence, two children in the same grade, say one starting the training in August, and one half a year later, start with different amounts of schooling. This implies that person-level random intercept and slope are also influenced by the schooling and should hence be interpreted with some caution. Especially the random intercept is a baseline competence at the start of the training here. In another application, time zero could represent the beginning of the school year. Alternatively, an additional (smooth) term could be added to the model, representing the amount of schooling received, potentially also for the slope. Given enough variation in the start time of the training, the two effects can be disentangled.

Note, that the conditional independence assumption means that beyond current ability, no other factors govern test behavior. Especially, repeated measurement with the same items has to proceed with ample time between measurement occasions so that memory effects do not violate conditional independence. Items need to be drawn without replacement at the same measurement occasion to rule out local item dependence.

In a correctly specified SGLDRM, which includes a correct specification of the spline and the random effects, all model parameters are consistently estimated. The larger the sample, the higher the precision of the estimation of the true global course of learning. When the random effects structure is misspecified, the model parameters still converge with growing sample size, but approach the value minimizing the Kullback-Leibler information (Heagerty & Kurland, 2001). In other words, the random effects specification, that is, the "linear deviation" part of the SGLDRM, needs to be checked. The random effects need to be modified if necessary, for example, by adding a non-smooth ability jump or drop, say after a summer break.

Assuming many measurement occasions per person, the SGLDRM can be extended to allow for more complex deviations from the global course: Introducing latent residuals, that is, $\vartheta_{jt} = \lambda(t) + \delta_j + \gamma_j t + \zeta_{jt}$ for stochastically independent ζ_{jt}, helps to study whether linear deviations are appropriate. The variance of the residuals, as well as their (mean) direction are informative. Quadratic or higher order polynomial terms could be added (or even person-level splines for the deviations). We caution that naive model checking of an SGLDRM's deviance residuals could be misleading, since binary data leads to clustered deviance residuals.

In the context of the dynamic measurement model (DMM) framework (Dumas & McNeish, 2017), trajectories are seen as nested within students who are nested within schools (Dumas et al., 2020). The GAMM framework is flexible enough to incorporate random effects for the school level with a specification parallel to that of Dumas et al. (2020). Similarly, also exponential growth curves which were found to be appropriate for a large data set of math development (Dumas et al.,

2020) are an alternative to linear deviations. Such extensions are beyond the scope of the current work, as they require a study of model identification, the parametrization of the deviations needs to be interpretable to be useful beyond graphical inspection. As dynamic IRT models are actively researched (Dumas et al., 2020), future work should also aim to include local item dependence due to repeated administration of the very same item.

6.2.3 Interpreting Correlations of Time Series

Correlations of time series are often positive if a global trend underlies all series. This has been frequently observed in the domain of macroeconomic time series (e.g., Gilbert & Meijer, 2005). In the context of educational time series, a positive global CoL and a naive calculation of correlations might suggest similarity of CoLs. Especially when (mean) time series are used to compare two or more groups, positive correlations might be spurious results of a global trend and hence potentially misleading.

The person-level deviations from the global CoL in the SGLDRM do not have this problem. They are measures of individual differences in CoL free from the spurious correlations. Once the model is calibrated, the CoL of new persons might be tracked. Then person-level slopes from the SGLDRM might be an indicator of intervention effects, even in single case research designs. Similarly, group comparisons based on group-specific deviations from the global CoL are a potential model extension addressing the issue. For this, a group mean CoL needs to be added to the model equation, represented, say, by a linear function in time.

6.2.4 Null and Perfect Scores

Maximum likelihood (ML) estimation of person ability is often not recommended, since perfect scores or null scores are a problem for the method. Here, however, it is unlikely that a person would have perfect or null scores at all measurement occasions. We recommend to remove persons with all null or all perfect scores from the data set. In both cases, the test at hand is inappropriate to track the CoL.

In case of very short assessments and very few measurement occasions, perfect and null scores might appear though the test is appropriate. The analyst might prefer a Bayesian strategy instead, for example, employ maximum a posteriori (MAP) or expected a posteriori (EAP) estimation of the two deviation factors. Note, that the likelihood ratio test breaks down, as it requires ML estimates. Alternatives for this situation include Bayes factors (Berger & Delampady, 1987).

6.2.5 Person-Level LRT Power

Our simulation results show that statistical power to detect deviations from the global course is low if the individual and global course intersect. Being able to identify students starting above average and falling behind is important in applications, but time spent testing could be less valuable than regular schooling. However, longer baseline assessments are practically feasible in many contexts and reduce uncertainty in the intercept. This in turn improves slope estimation (in analogy to simple linear regression; Atkinson & Donev, 1992, p. 38ff.) and increases power of the LRT.

6.2.6 Interpretation in Application

Due to the lack of empirical values when it comes to putting the SGLDRM into practice there is no ideal guideline for teachers to interpret the outcomes yet. However, well-thought-out guidelines are essential in order for diagnostics to be interpreted correctly by non-experts (cf. Zeuch et al., 2017). There are roughly three ways to present SGLDRM results: The global trend as well as the linear deviation can be presented graphically, facilitating a visual assessment of a student's as well as the general development. Apart from that, the SGLDRM gives three numerical scores per student for interpretation and diagnosis – an intercept, a slope, and whether or not the deviations from the global curve are (statistically) significant. In terms of the decision and judgment inferences of Hopster-den Otter et al. (2019), it would be important to provide guidelines for interpreting each outcome correctly. For better accessibility the parameter values could further be simplified by color schemes, graphical methods, and written explanations. However, one should always keep in mind, that a test performance should be interpreted in relation to other student performances and the teacher's assessments in order to meet the generalization inference (Hopster-den Otter et al., 2019). When implemented and communicated properly the SGDLRM offers valuable information on student development and can support teachers in educational diagnostics.

References

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*(1), 3–16. https://doi.org/10.1007/BF02294143

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: Population parameter estimation. *Journal of Multivariate Analysis*, *95*(1), 1–22. https://doi.org/10.1016/j.jmva.2004.07.005

Atkinson, A. C., & Donev, A. N. (1992). *Optimum experimental designs*. Clarendon.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science, 2*(3), 317–335. https://doi.org/10.1214/ss/1177013238

Binder, H., & Tutz, G. (2008). A comparison of methods for the fitting of generalized additive models. *Statistics and Computing*, *18*(1), 87–99. https://doi.org/10.1007/s11222-007-9040-0

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25. https://doi.org/10.1080/01621459.1993.10594284

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury.

Cho, S.-J., Athay, M., & Preacher, K. J. (2013). Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 353–381. https://doi.org/10.1111/j.2044-8317.2012.02058.x

Cho, S.-J., Brown-Schmidt, S., De Boeck, P., & Naveiras, M. (2022). Space-time modeling of intensive binary time series eye-tracking data using a generalized additive logistic regression model. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000444

Cho, S.-J., Preacher, K. J., Yaremych, H. E., Naveiras, M., Fuchs, D., & Fuchs, L. S. (2022). Modelling multilevel nonlinear treatment-by-covariate interactions in cluster randomized controlled trials using a generalized additive mixed model. *British Journal of Mathematical and Statistical Psychology*. Advance online publication. https://doi.org/10.1111/bmsp.12265

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

Debeer, D., Janssen, R., Buchholz, J., & Hartig, J. (2014, July 22–25). *Modeling differences in item-position effects in the PISA 2009 reading assessment within and between schools* [Conference presentation]. 79th International Meeting of the Psychometric Society. Madison, WI, United States.

Doebler, A., Doebler, P., & Holling, H. (2013). Optimal and most exact confidence intervals for person parameters in item response theory models. *Psychometrika*, *78*(1), 98–115. https://doi.org/10.1007/s11336-012-9290-4

Dumas, D., McNeish, D., & Greene, J. A. (2020). Dynamic measurement: A theoretical-psychometric paradigm for modern educational psychology. *Educational Psychologist*, *55*(2), 88–105. https://doi.org/10.1080/00461520.2020.1744150

Dumas, D. G., & McNeish, D. M. (2017). Dynamic measurement modeling: Using nonlinear growth models to estimate student learning capacity. *Educational Researcher*, *46*(6), 284–292. https://doi.org/10.3102/0013189X17725747

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495–515. https://doi.org/10.1007/BF02294487

Embretson, S. E. (1997). Structured ability models in tests designed from cognitive theory. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 223–236). Ablex.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica*, *37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. de Gruijter & L. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97–110). Wiley.

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika*, *54*(4), 599–624. https://doi.org/10.1007/BF02296399

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal*, *21*(2), 449–460. https://doi.org/10.3102/00028312021002449

Gilbert, P. D., & Meijer, E. (2005). *Time series factor analysis with an application to measuring money* (Research Report 05F10). University of Groningen, Research School SOM.

Green, P., & Silverman, B. (1994). *Nonparametric regression and generalized linear models: A roughness penalty approach*. Chapman & Hall. https://doi.org/10.1007/978-1-4899-4473-3

Gu, C., & Kim, Y. (2002). Penalized likelihood regression: General formulation and efficient approximation. *Canadian Journal of Statistics*, *30*(4), 619–628. https://doi.org/10.2307/3316100

Heagerty, P. J., & Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, *88*(4), 973–985. https://doi.org/10.1093/biomet/88.4.973

Hecht, M., Hardt, K., Driver, C. C., & Voelkle, M. C. (2019). Bayesian continuous-time Rasch models. *Psychological Methods*, *24*(4), 516–537. https://doi.org/10.1037/met0000205

Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2019). A general framework for the validation of embedded formative assessment. *Journal of Educational Measurement*, *56*(4), 715–732. https://doi.org/10.1111/jedm.12234

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Klauer, K. C. (1991). Exact and best confidence intervals for the ability parameter of the Rasch model. *Psychometrika*, *56*(3), 535–547. https://doi.org/10.1007/BF02294489

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813–1824. https://doi.org/10.1016/j.compedu.2011.02.003

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-4310-4

Kuhn, J.-T., Schwenk, C., Raddatz, J., Dobel, C., & Holling, H. (2018). *CODY-M 2-4: CODY-Mathetest für die 2.-4. Klasse* [CODY-M 2-4: CODY-math-test for 2.-4. Grade]. kaasa health.

Le, L. T. (2007, July 9–13). *Effects of item positions on their difficulty and discrimination: A study in PISA science data across test language and countries* [Conference presentation]. 72nd International Meeting of the Psychometric Society, Tokyo, Japan.

Lenhard, A., Lenhard, W., & Gary, S. (2019). Continuous norming of psychometric tests: A simulation study of parametric and semi-parametric approaches. *PloS ONE*, *14*(9), Article e0222279. https://doi.org/10.1371/journal.pone.0222279

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*(2), 233–245. https://doi.org/10.1007/BF02294018

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*(1), 5–9. https://doi.org/10.1111/j.1750-8606.2009.00109.x

Mühling, A., Gebhardt, M., & Diehl, K. (2017). Formative Diagnostik durch die Onlineplattform LEVUMI [Formative diagnostics through the online platform LEVUMI]. *Informatik Spektrum*, *40*(6), 556–561. https://doi.org/10.1007/s00287-017-1069-7

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*(2), 165–187.

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. Oxford University Press.

Powell, M. J. D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives* (Numerical Analysis Report, DAMTP 2009/ NA06). Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment*, *84*(3), 228–238. https://doi.org/10.1207/s15327752jpa8403_02

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, *6*(1), 15–32. https://doi.org/10.1214/ss/1177011926

Rost, J., & Spada, H. (1983). Die Quantifizierung von Lerneffekten anhand von Testdaten [Quantifying learning effects on the basis of test data]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *4*(1), 29–49.

Saha, A. (2016). *Bayesian analysis of item response theory and its application to longitudinal education data* (Doctoral dissertation, University of Connecticut).

Schurig, M., Jungjohann, J., & Gebhardt, M. (2019). *Handbuch für Lehrkräfte im Anwendungsbereich Verhalten und Empfinden – Lern-Verlaufs-Monitoring Levumi* [Manual for teachers in the field of application behavior and sensation – learning progress monitoring Levumi]. TU Dortmund University. https://doi.org/10.17877/DE290R-20376

Schwenk, C., Kuhn, J.-T., Gühne, D., Doebler, P., & Holling, H. (2017). Auf Goldmünzenjagd: Psychometrische Kennwerte verschiedener Scoringansätze bei computergestützter Lernverlaufsdiagnostik im Bereich Mathematik [We are going on a gold coin hunt: Psychometric properties of different scorings in computer-based progress monitoring of mathematics ability]. *Empirische Sonderpädagogik, 9*, 123–142.

Souvignier, E., Förster, N., & Salaschek, M. (2014). quop: Ein Ansatz internetbasierter Lernverlaufsdiagnostik mit Testkonzepten für Mathematik und Lesen [quop: An internet-based approach to learning progress monitoring in math and reading]. In M. Hasselhorn, W. Schneider, & U. Trautwein (Eds.), *Lernverlaufsdiagnostik* (pp. 239–256). Hogrefe.

Strathmann, A. M., & Klauer, K. J. (2012). *LVD-M 2-4: Lernverlaufsdiagnostik-Mathematik für zweite bis vierte Klassen* [LVD-M 2-4: Diagnosing the course of learning in mathematics for second to fourth grades]. Hogrefe.

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Springer. https://doi.org/10.1007/0-306-47531-6

Wang, X., Berger, J. O., & Burdick, D. S. (2013). Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, *7*(1), 126–153. https://doi.org/10.1214/12-AOAS608

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Taylor & Francis. https://doi.org/10.1201/9781315370279

Wood, S. N., & Scheipl, F. (2017). *gamm4: Generalized additive mixed models using 'mgcv' and 'lme4'* [R package version 0.2-5]. https://CRAN.R-project.org/package=gamm4

Wu, Q., Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large-scale Assessments in Education*, *7*, Article 5. https://doi.org/10.1186/s40536-019-0073-6

Yoshida, T., & Naito, K. (2014). Asymptotics for penalised splines in generalised additive models. *Journal of Nonparametric Statistics*, *26*(2), 269–289. https://doi.org/10.1080/10485252.2014.899360

Zeuch, N., Förster, N., & Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessment: Results from tests and interviews. *Learning Disabilities Research & Practice*, *32*(1), 61–70. https://doi.org/10.1111/ldrp.12126

Appendix

1. Estimation Procedure

When using the statistical software R (R Core Team, 2020) to fit a GAMM, there are several functions to choose from, *mgcv* (Wood, 2017) and *gamm4* (Wood & Scheipl, 2017) being two packages to provide them. As mentioned in the main text thin plate regression splines are used as default by all of the functions. The gamm4 routines use the ML (maximum likelihood) method for parameter estimation in the case of a binomial family and *logit* as link function. A Laplace approximation to the (log-)likelihood is employed (Breslow & Clayton, 1993) and is maximized by bound optimization by quadratic approximation (BOBYQA) by default – a derivative free numerical optimization algorithm that iteratively minimizes a function by making use of quadratic models (Powell, 2009). The gamm routine of the mgcv package finds (approximate) ML estimators iteratively by penalized quasi likelihood (PQL) estimation. PQL estimation leads to asymptotic normality and efficiency of parameters estimates, including the spline (Yoshida & Naito, 2014).

While older simulation studies (Binder & Tutz, 2008) and current complex applications (Cho, Brown-Schmidt et al., 2022; Cho, Preacher et al., 2022) of GAMMs show that current estimation algorithms perform well when ample binary data is available, we have not investigated what the lowest feasible sample size for the SGLDRM is. We expect that it is hard to give general recommendations, as performance will depend on factors like number of items, number of measurement occasions, as well as distributions of latent variables and item difficulties. In case the SGLDRM is to be employed for data of less than 1000 individuals, we recommend targeted parameter recovery simulations to check for bias and RMSE of structural model parameters.

2. Latent Variance

While the model for the latent ability mean is very flexible, the time-specific latent variance is constrained by the model specification. One implication of the linear growth is that the variance of the person parameters will increase with time when the variance of the slopes is substantial: The following formula can be interpreted as indicating that the individual deviations from the global course spread the ability spectrum, though a negative correlation of the slopes and intercepts might delay this effect or reverse it initially when observation time is limited. More specifically, let ϑ_t denote a random variable for the marginal distribution of ability at time t, then

$$\mathrm{Var}(\vartheta_t) = \sigma_\delta^2 + 2t\rho\sigma_\delta\sigma_\gamma + t^2\sigma_\gamma^2. \tag{15}$$

Hence, depending on the entries of Σ, the variance of ϑ_t can increase or decrease with t, but is quadratic in t. Similarly, covariances are highly structured. The following model-implied formula for the SGLDRM's latent covariances can be used for model checking by calculating the covariance for two time points with the help of a different model, say a bivariate Rasch model. For two time points t_1 and t_2 the covariance of the latent variables is given by

$$\mathrm{Cov}(\vartheta_{t_1}, \vartheta_{t_2}) = \mathrm{Cov}(\delta_j + \gamma_j t_1, \delta_j + \gamma_j t_2) = \sigma_\delta^2 + (t_1 + t_2)\rho\sigma_\delta\sigma_\gamma + t_1 t_2 \sigma_\gamma^2 \tag{16}$$

which allows for an increase or decrease of the covariance but which is bilinear in t_1 and t_2.