List *et al. Large-scale Assess Educ (2017) 5:15*
DOI 10.1186/s40536-017-0049-3

Large-scale Assessments in Education

RESEARCH Open Access

Performance decline in low-stakes educational assessments: different mixture modeling approaches

Marit K. List[1*], Alexander Robitzsch[1,2], Oliver Lüdtke[1,2], Olaf Köller[1,2] and Gabriel Nagy[1]

*Correspondence: list@ipn.uni-kiel.de
[1] Leibniz Institute for Science and Mathematics Education, Kiel, Germany
Full list of author information is available at the end of the article

Abstract

Background: In low-stakes educational assessments, test takers might show a performance decline (PD) on end-of-test items. PD is a concern in educational assessments, especially when groups of students are to be compared on the proficiency variable because item responses gathered in the groups could be differently affected by PD. In order to account for PD, mixture item response theory (IRT) models have been proposed in the literature.

Methods: In this article, multigroup extensions of three existing mixture models that assess PD are compared. The models were applied to the mathematics test in a large-scale study targeting school track differences in proficiency.

Results: Despite the differences in the specification of PD, all three models showed rather similar item parameter estimates that were, however, different from the estimates given by a standard two parameter IRT model. In addition, all models indicated that the amount of PD differed between tracks, in that school track differences in proficiency were slightly reduced when PD was accounted for. Nevertheless, the models gave different estimates of the proportion of students showing PD, and differed somewhat from each other in the adjustment of proficiency scores for PD.

Conclusions: Multigroup mixture models can be used to study how PD interacts with proficiency and other variables to provide a better understanding of the mechanisms behind PD. Differences between the presented models with regard to their assumptions about the relationship between PD and item responses are discussed.

Keywords: Educational assessments, Mixture IRT models, Performance decline, Group comparisons, Aberrant response behavior

Background

One main purpose of large-scale assessments (LSAs) is to provide policy-makers and educational institutions with information about students' proficiency. Having no personal consequences, educational assessments are low-stakes tests for the test takers (Baumert and Demmrich 2001; DeMars 2000; Penk et al. 2014; Wise and DeMars 2005) and, therefore, some test takers might not invest full effort throughout the test, resulting in an underestimation of their true proficiency levels. As a consequence, the estimates of proficiency scores and the inferences based on them (e.g., group differences) are likely to be biased (Eklöf 2010; Wise and DeMars 2005; Wise and Kong 2005).

A common observation in educational LSAs is that the probability of a test taker giving a correct response decreases for items at the end of the test (Wu 2010). On the test taker's side, this can be viewed as a performance decline (PD), which can be considered as a type of *aberrant response behavior* reflected in unexpectedly high rates of incorrect or omitted responses for end-of-test items (Cao and Stokes 2008; Schnipke and Scrams 1997; Suh et al. 2012). Whereas, in high-stakes tests, aberrant response behavior on end-of-test items is often attributed to test speededness (Bolt et al. 2002), in low-stakes tests, it is assumed that aberrant response behavior is related to a decline in test-taking effort (Wise and Kong 2005). In order to explore PD in educational assessments, mixed-effects models (De Boeck and Wilson 2004) have been proposed to analyze item position effects by comparing the probabilities of a correct response when the item is presented in different positions (Debeer and Janssen 2013). An alternative strategy proposes using mixture models with categorical latent variables in order to identify the latent classes of test takers who differ in their test-taking behavior (Mislevy and Verhelst 1990; Rost 1990). Among others, Bolt et al. (2002), Yamamoto (1995), and Jin and Wang (2014) introduced mixture models to separate test takers who show aberrant response behavior that corresponds to PD from test takers who respond to all items with full effort.

The aim of our study was to examine the utility of three mixture models to handle PD in low-stakes tests and to explore the differences and similarities in the conclusions drawn from these models. In this study, we applied the mixture models to investigate PD in the mathematics test of a German LSA. First, we explored the differences between the PD of test takers attending different school types. Next, we investigated whether PD affects the estimation of group differences in proficiency. Then, we compared existing mixture models with regard to the differences and similarities in their estimation of PD. Furthermore, we demonstrate how to fit these models using standard software such as M*plus* (Muthén and Muthén 1998–2012).

This article is organized as follows. First, we will provide an overview of the research on PD in educational assessments. We will then proceed to present three mixture models for PD that were extended to multigroup settings. After that, we will apply these models to a low-stakes mathematics test in order to compare their performance and parameter estimates. Finally, we will present and discuss the differences and similarities of these models regarding their measurement of PD and estimated group differences in proficiency.

Performance decline in educational assessments

If PD is present, the probability of providing a correct response does not depend solely on item parameters and a person's proficiency. Simulation studies have shown that parameter estimates of end-of-test items are biased if PD is not taken into account (Oshima 1994; Suh et al. 2012). This is of particular concern when parameter estimates of items are going to be treated as known in other applications, for example, in adaptive testing, as part of a calibrated item pool, or for test construction purposes (Davey and Lee 2011; van Barneveld 2007). In addition, test scores obtained under PD conditions are likely to be lower than they would be if the test taker had invested full effort throughout the test. Thus, when PD is ignored, proficiency scores are underestimated—the stronger

the PD effects are and the more test takers experience PD, the stronger the underestimation of the sample's average proficiency will be.

Several authors have studied associations between PD and test takers' characteristics, such as ethnicity, language skills, or gender. Bolt et al. (2002) analyzed a high-stakes college placement test for mathematics and found that the number of students showing PD differed between different ethnic groups. Yamamoto and Everson (1995) assessed PD in a high-stakes reading comprehension test for university students. They also found differences between the PD of different ethnic groups. Furthermore, they found that nonprimary-language speakers showed an earlier onset of PD than primary-language speakers. For a high-stakes reasoning test, Schnipke and Pashley (1997) also found that nonprimary-language speakers were more strongly affected by PD than primary-language speakers.

For low-stakes tests, differences in PD have been reported on a country level for the PISA assessments (Debeer et al. 2014; Hartig and Buchholz 2012; Jin and Wang 2014). Nagy et al. (2016) also found PD effects to differ between school types in the German PISA 2012 study. Furthermore, male test takers have been found to show more guessing behavior and lower levels of test-taking motivation than female test takers (DeMars et al. 2013).

Due to the differential effects found, these results imply that ignoring PD might affect the estimations of average proficiency levels in a group-specific way. Thus, when investigating group differences in proficiency, the extent to which the differences found might be caused by differences in PD instead of true differences in proficiency should be explored (DeMars et al. 2013; Denis and Gilbert 2012; Mittelhaëuser et al. 2015).

Mixture models of performance decline

Mixture *item response theory* (IRT; e.g., Embretson and Reise 2000) models can be used to identify test takers showing PD. We will refer to these models as *mixture PD models* in the rest of this article. In general, mixture models assume that the population consists of subgroups, for which the model parameters differ in particular ways (Mislevy and Verhelst 1990; Rost 1990). These subgroups are also referred to as *latent classes* since they are not observed. Based on their individual responses, the probabilities for each test taker of belonging to one of these latent classes are estimated along with the other model parameters. Mixture PD models consist of two or more latent classes, where one class represents test-taking behavior that reflects full effort throughout the test (the *no-decline class*), and the other classes represent test-taking behavior that reflects PD (the *decline classes*).

Of these, two models, the two-class mixture model of Bolt et al. (2002) and the HYBRID model of Yamamoto (1995) have been applied to several empirical and simulated data sets (e.g., Boughton and Yamamoto 2007; Cao and Stokes 2008; Hailey et al. 2012; Mittelhaëuser et al. 2013; Mittelhaëuser et al. 2015; Suh et al. 2012; Wollack et al. 2003; Yamamoto and Everson 1997). Bolt et al. (2002) developed their two-class mixture model to reduce the bias in item parameters that is caused by test speededness in high-stakes tests. The aim of the model is to identify the group of test takers who run out of time and show PD (i.e., the decline class). PD is reflected in higher difficulty parameters for end-of-test items. The item parameters of the no-decline class are expected to be

unbiased. Later on, the model was also used to model differences in test-taking motivation in low-stakes tests in order to separate motivated from unmotivated test-taking behavior (e.g., Mittelhaëuser et al. 2015).

The HYBRID model of Yamamoto (1995) was initially developed to model changes in test-taking behavior that occur under conditions of test speededness in high-stakes tests. The HYBRID model focuses on determining the position in which a test taker switches from effortful response behavior to aberrant response behavior when running out of time. The item responses that reflect aberrant response behavior are ignored for the estimation of proficiency and item parameters; therefore, item parameter and proficiency estimates are expected to be unbiased. The HYBRID model has also been used to investigate PD in low-stakes tests (Cao and Stokes 2008).

Recently, Jin and Wang (2014) proposed a multiclass mixture model to account for PD. Similar to Yamamoto (1995) HYBRID model, it makes it possible to determine the point where test takers alter their test-taking behavior. In contrast to the HYBRID model, Jin and Wang (2014) do not assume that test takers switch to entirely aberrant response behavior that is unrelated to proficiency. Rather, the authors model PD as a decline in test performance so that the probability of providing a correct response decreases after the onset of PD.

In a recent study, the HYBRID model (Yamamoto 1995) and the two-class mixture model of Bolt et al. (2002) were compared with regard to their capability to reduce the bias in item parameters that is caused by test speededness (Suh et al. 2012). Both models were found to be equally capable of estimating the true item parameters and outperformed a standard IRT model that did not account for test speededness. In addition, in both cases, the results were not affected by the coding of the items not reached by the examinees (i.e., missing vs. incorrect). However, the focus of the study of Suh et al. (2012) was bias in item parameters; the effects of PD on the estimation of proficiency therefore need further investigation. Furthermore, the model of Jin and Wang (2014) has not yet been investigated thoroughly, and its performance has not yet been compared with that of the model of Yamamoto (1995) or the model of Bolt et al. (2002).

A general mixture IRT model of performance decline

In the three aforementioned mixture PD models, several common assumptions are made. It is assumed that test takers respond to the items in the order in which they are presented. It is further assumed that PD starts at some point in the test, that is, the test takers will have responded to at least the first item with full effort before their test-taking behavior reflects PD. Thus, a *switching point* (von Davier and Yamamoto 2007) exists at which decline onset can be observed, dividing the item response vector of a test taker into two parts: in the first part, item responses depend solely on the test taker's proficiency and on the item characteristics. In the second part, item responses depend on a latent person variable underlying the aberrant response behavior that reflects PD. Depending on the definition of PD as specified in the particular model, aberrant response behavior may or may not be related to proficiency. For both parts, the relationship between item responses and the latent person variable can be modeled within the IRT framework. We assume that, before the switching point, a two-parameter logistic model (2PLM; Birnbaum 1968) holds. However, after the switching point, we assume

that item and person parameters might be different. Furthermore, we assume that some test takers show full effort throughout the test. Thus, the sample consists of a mixture of test takers with and without PD.

Based on these assumptions, a *general mixture PD model* can be described, which contains the aforementioned models as special cases. In the following, we describe such a model for the multigroup case, although the framework can be easily adapted for single group situations. According to the 2PLM, the probability of person p providing a correct response to item i depends on the person's proficiency (θ_p) and the item's slope and intercept parameter (α_i,β_i). The person's individual switching point (δ_p) equals the last item position before decline onset. Thus, the switching point can take values $\delta_p = 1, \ldots, I$, where I is the total number of items in the test. If $\delta_p = i$, PD will start after item i. If $\delta_p = 1$, PD will start after the first item. If $\delta_p = I$, PD will not occur at all in the test.

Let η_{pi} be the logit of the probability of correct response of person p to item i: $\eta_{pi} = \text{logit}\, P\left(X_{pi} = 1|\theta_p, \delta_p\right)$. For the general mixture PD model, the conditional item response probabilities are defined as

$$\eta_{pi} = \begin{cases} \alpha_i \cdot \theta_p + \beta_i, & \text{if } \delta_p \leq i, \\ \tilde{\alpha}_{ig} \cdot \tilde{\theta}_p + \tilde{\beta}_{ig}, & \text{if } \delta_p > i, \end{cases} \tag{1}$$

where θ_p denotes the proficiency variable, α_i is the slope parameter or item discrimination and β_i is the intercept parameter[1], both before the switching point, that is, for $\delta_p \leq i$. Note that the item parameters applied to responses before the switching point contain no group index g, so that we assume them to be invariant across groups, although this assumption can be relaxed.

The parameters $\tilde{\alpha}_{ig}$ and $\tilde{\beta}_{ig}$ denote the slope and intercept parameter in group g after the switching point, that is, for $\delta_p > i$, respectively, and $\tilde{\theta}_p$ denotes the person variable underlying aberrant response behavior. When a test taker shows PD, the proficiency variable θ_p can change to $\tilde{\theta}_p$, which would no longer be interpreted as proficiency. Equation 1 is a mixture 2PLM where the latent classes capture the magnitude of the individual PD. If $\delta_p = I$ for all test takers in the sample, Eq. 1 is reduced to $\eta_{pi} = \alpha_i \cdot \theta_p + \beta_i$, which is the standard 2PLM.

The joint distribution of θ and δ is defined based on the assumption of how proficiency and switching point are related. Typically, θ is considered to be a normally distributed continuous variable. By definition, δ is considered to be a discrete variable. One way of specifying the joint distribution $P\left(\theta, \delta|G = g\right)$ is to consider the conditional distribution of θ with respect to δ in group g (i.e., $\delta|G = g$), that is, $P(\theta,\delta|G = g) = P(\theta|\delta, G = g) \cdot P(\delta|G = g)$. Within each class and each group, a normal distribution of θ is assumed, that is, $P\left(\theta|\delta = i, G = g\right) = N\left(\mu_{ig}, \sigma^2_{ig}\right)$. Typically, in the case of many classes, restrictions are imposed on the means μ_{ig} and standard deviations σ_{ig} (e.g., Yamamoto and Everson 1997).

The dependencies between θ and δ as assessed by multigroup mixture PD models could be of substantive interest in real applications. For example, research suggests that

[1] Note that the item difficulty b_i is the negative of the ratio of intercept to slope parameter (Hambleton et al. 1991), that is, $b_i = -\beta_i/\alpha_i$.

students with lower proficiency estimates are more likely to show aberrant response behavior, as reflected in PD (Bolt et al. 2002; De Boeck et al. 2011; Wise and Kong 2005); mixture PD models make it possible to examine such hypotheses. In addition, researchers might expect that the strengths of the relationship between θ and δ differ between groups (e.g., groups assessed in low-stakes vs. high-stakes conditions). Again, the multigroup setup of mixture PD models allows such hypotheses to be explicitly tested.

In addition, the multigroup mixture PD models provide estimates for the proportions of decline classes in group g, $\pi_{ig} = P(\delta = i|G = g)$ for $i < I$, including the proportions of test takers not affected by PD, $\pi_{Ig} = P(\delta = I|G = g)$, thereby allowing researchers to examine whether the proportion of test takers not affected by PD differs between groups, and to investigate group differences in the prevalence of earlier and later onsets of PD.

However, the assessment of latent class probabilities, π_{ig}, might not succeed in mixture PD models that contain many latent classes without imposing additional constraints. In order to solve this problem, Cao and Stokes (2008) proposed a cumulative probability function of switching points, which was also employed by Jin and Wang (2014) in their mixture PD model. In the multigroup case, the function recurs on the probability of the no-decline class in group g, π_{Ig}, and the probabilities of the decline classes, π_{ig}, for $i < I$. Cao and Stokes (2008) assume that the cumulative probability function of δ_p depends on a shape parameter $\omega > 0$, so that

$$P\left(\delta_p \leq i|G = g\right) = \frac{i^{\omega_g}}{(I-1)^{\omega_g}}, \quad \text{for } i < I. \tag{2}$$

The shape parameter ω_g defines the form of the cumulative probability curve in group g: if $\omega_g = 1$, the probability increases linearly, if $\omega_g > 1$, the increase is convex, and if $\omega_g < 1$, the increase is concave. With Eq. 2, the curve of π_{ig} is described as

$$\pi_{ig} = \frac{i^{\omega_g} - (i-1)^{\omega_g}}{(I-1)^{\omega_g}} \cdot \left(1 - \pi_{Ig}\right), \quad \text{for } i < I. \tag{3}$$

In Eq. 3, only two parameters are estimated per group, the proportion of the no-decline class π_{Ig} and the shape parameter ω_g; this reduces the number of parameters for long tests with many switching points. Because of its parsimony, we used Eq. 3 in our study to model the probability distributions for the decline classes in mixture PD models with many classes. However, other functions could be used as well.

The cumulative probability function of δ_p (Eq. 2) used in our application allows researchers to explicitly test hypotheses about group differences in the proportion of examinees affected by PD via the estimates of π_{Ig} (Eq. 3). In addition, the function allows for a straightforward comparison of the ω_g-parameters that indicate whether the onset of PD occurs later (higher values of ω) or earlier in the test (lower values of ω). However, as the group-specific proportions of PD onset points π_{ig} is a rather complex function of ω_g and π_{Ig} (Eq. 3), we suggest using graphical aids for comparing the distribution of onset points across groups.

Finally, it should be noted that the multigroup mixture PD models make it possible to assess not only the group-specific distributions of the onset points of PD but also the magnitude of PD in each group. However, the magnitude of PD depends on the specific restrictions imposed on the general mixture PD model (Eq. 1): in some models, the

magnitude of PD is expressed via the $\tilde{\beta}_{ig}$-parameters (Bolt et al. 2002; Yamamoto 1995), whereas, in others, the magnitude of PD is quantified by the person variable $\tilde{\theta}_p$ (Jin and Wang 2014).

Multigroup mixture IRT models of performance decline

By placing specific parameter restrictions, the general mixture PD model can be transformed into multigroup versions of each of the three aforementioned mixture PD models (Bolt et al. 2002; Jin and Wang 2014; Yamamoto 1995). Since PD is modeled differently in the specific mixture PD models, group comparisons for each mixture PD model involve different parameters. In the following, we will describe how the general mixture PD model can be extended to realize the mixture PD models of Bolt et al. (2002), Yamamoto (1995), and Jin and Wang (2014), allowing for multigroup comparisons.

The two-class mixture model of Bolt et al. (2002)

Bolt et al. (2002) proposed a mixture PD model with two latent classes (which will further be referred to as the 2PDM): while test takers in the no-decline class do not show PD, test takers in the decline class show lower performance on end-of-test items. Therefore, in the decline class, items appear to be more difficult than in the no-decline class, resulting in lower item intercept parameter estimates after the switching point. In their original formulation of the 2PDM, Bolt et al. (2002) suggested specifying the switching point i_0 in advance (see also De Boeck et al. 2011; Wollack et al. 2003), so that it refers to an item position up to which responses are expected to not be affected by PD. When put into the multigroup context, this specification implies that, in each group, the switching point is a dichotomous variable that can take two values, that is, $\delta_p = i_0$ for all test takers showing PD, and $\delta_p = I$ for all test takers not showing PD. Accordingly, in each group, there are two latent classes with proportions $\pi_{i_0 g} = P\left(\delta = i_0 | G = g\right)$ for the decline class and $\pi_{Ig} = P(\delta = I | G = g)$ for the no-decline class.

The model was first proposed as a mixture Rasch model with equal item discrimination parameters for all items. Recent applications have also considered extensions based on a 2PLM (Cao and Stokes 2008; Suh et al. 2012), which we also used in our study. For each group, we restricted the item discriminations to be equal in both latent classes (i.e., $\tilde{\alpha}_{ig} = \alpha_i$; see Cao and Stokes 2008). After the switching point, item intercept parameters were allowed to change but item responses to still depend solely on the item parameters and proficiency. Thus, the person variable underlying the aberrant response behavior would still be regarded as proficiency, hence $\tilde{\theta}_p = \theta_p$. Within each group, item intercept parameters were constrained to be lower in the decline class (i.e., $\tilde{\beta}_{ig} \leq \beta_i$ for each g). Equation 1 can be altered to realize the multigroup version of the 2PDM with

$$\eta_{pi} = \begin{cases} \alpha_i \cdot \theta_p + \beta_i, & \text{if } \delta_p \leq i_0, \\ \alpha_i \cdot \theta_p + \tilde{\beta}_{ig}, & \text{if } \delta_p > i_0. \end{cases} \tag{4}$$

The mean proficiency is allowed to vary across groups and classes, while the standard deviation is held equal across latent classes, that is, $\theta \sim N(\mu_{ig}, \sigma_g^2)$. We assumed that groups could differ in the intercept parameters after the switching point, $\tilde{\beta}_{ig}$ (see Eq. 4). In addition, we did not impose any constraints on the latent class proportions, $\pi_{i_0 g}$ and π_{Ig}.

Note that, in the common specification of the 2PDM, the (arbitrarily) specified switching point i_0 does not identify the item position in which PD first occurs. Rather, researchers are required to compare the estimates of β_i and $\tilde{\beta}_{ig}$ and to identify the position from which these estimates show meaningful deviations from each other. An alternative is to identify the point at which PD first occurs by comparing the model-data-fit of the 2PDM, specified with different values of i_0, and selecting the switching point that provides the best fit to the data at hand. This procedure has the advantage that many β_i-parameters that would otherwise be specified to be class- and group-specific (i.e., $\tilde{\beta}_{ig}$-parameters) are estimated on the basis of a larger number of item responses. The procedure can be extended to identify group-specific switching points, but the merits of such an approach seem to be limited. When a common switching point is assumed, researchers can inspect the results for group differences in the onset of PD by comparing the estimates of $\tilde{\beta}_{ig}$ with the corresponding estimates of β_i. Results showing negligible discrepancies between $\tilde{\beta}_{ig}$- and β_i-parameters after the switching point i_0 indicate a later onset point of PD in this group.

Hence, the multigroup 2PDM makes it possible to (1) assess the relationship between proficiency and PD on the basis of the latent class means μ_{ig}, (2) estimate the proportion of examinees not affected by PD in each group (π_{Ig}), and (3) quantify the strengths of PD in each group by examining the discrepancies between $\tilde{\beta}_{ig}$- and β_i-parameters. In addition, the latter parameters also make it possible to (4) inspect the results for group differences in the onset points of PD.

The HYBRID model of Yamamoto (1995)

In his HYBRID model (further referred to as the HYBRID), Yamamoto (1995) assumes that, after the switching point, item responses no longer reflect test takers' proficiency. Instead, the probability of a test taker providing a correct response to items after PD onset is independent of proficiency; rather, it corresponds to an item-specific response threshold. While the model was originally proposed to model random guessing in high-stakes tests under speededness conditions, it can also be applied to other types of aberrant response behavior reflecting PD (e.g., Suh et al. 2012).

PD onset can occur in all item positions except the first one, hence, there are multiple decline classes, one for each item position, so that $\delta_p = 1, \ldots, I$ in each group. Within each group, the response thresholds can be specified to be either item-specific or equal for all items. The latter assumption is reasonable in cases where all items share a similar response format, such as multiple-choice items with the same number of options.

Since item responses after the switching point do not depend on proficiency, the slope parameter after the switching point is set to zero in each group ($\tilde{\alpha}_{ig} = 0$, for each g). Since the slope parameters after the switching point are set to zero, the item responses do not depend on the person variable, $\tilde{\theta}_p$, meaning that the person variable underlying aberrant response behavior is not defined. To keep the model identified, we set $\tilde{\theta}_p = \theta_p$. Note that $\tilde{\theta}_p$ can be fixed to any other value, as it does not impact on η_{pi} after the switching point. The intercept parameter after the switching point is constrained to a common response threshold within each group, which we restricted to be the same for all items in all decline classes, that is, $\tilde{\beta}_{ig} = \tilde{\beta}_g$ for all i.

Thus, within each group, the probability of a correct response after PD onset is the same for all test takers and items regardless of proficiency (or any other person variable underlying aberrant response behavior) and regardless of the location of PD onset. Based on these specifications, the multigroup HYBRID can be derived by altering Eq. 1 to

$$\eta_{pi} = \begin{cases} \alpha_i \cdot \theta_p + \beta_i, & \text{if } \delta_p \leq i, \\ 0 \cdot \theta_p + \tilde{\beta}_g, & \text{if } \delta_p > i. \end{cases} \tag{5}$$

The distribution of decline class probabilities within each group was assumed to follow the function described in Eqs. 2 and 3. We assumed that the mean of θ in each PD class assessed in each group is a function of δ, while the standard deviation of θ is the same for all latent classes but differs between groups. More specifically, we modeled the mean of the conditional θ distributions across latent classes as a linear function (Yamamoto and Everson 1997) with the mean of the no-decline class, μ_{Ig}, as the intercept and a group-specific slope parameter ρ_g:

$$\mu_{ig} = \mu_{Ig} + \rho_g \cdot (I - i). \tag{6}$$

The value of ρ_g in Eq. 5 shows how θ and δ are related: if ρ_g is negative, average proficiency is lower, the earlier the PD onset. If ρ_g is positive, average proficiency would be lower for later PD onsets.

Taken together, the multigroup HYBRID makes it possible to assess the group-specific (1) relationships between proficiency and PD via the parameters ρ_g, (2) proportions of examinees not affected by PD by means of the latent class proportions π_{Ig}, and (3) strengths of PD effects that are governed by the response thresholds $\tilde{\beta}_g$. Compared to the multigroup 2PDM, the HYBRID model presented here allows for (4) a more finely-grained assessment of the group-specific onset points of PD by inspecting the group-specific cumulative probability functions of δ_p (Eq. 2) and their parameters ω_g.

The multiclass mixture performance decline model of Jin and Wang (2014)

Jin and Wang (2014) proposed another multiclass mixture PD model (further referred to as the MPDM). Similar to the HYBRID, PD onset can occur after any item position throughout the test. In the MPDM, it is assumed that, after the switching point, θ_p is reduced according to a decrement function which depends on δ_p. Hence, when specified in a multigroup context, in the MPDM, PD is modeled via a group-specific change in the θ-variable instead of via changes in the intercept parameters as is the case in the 2PDM and the HYBRID model. More specifically, in the multigroup MPDM, it is assumed that, within each group, the value of the decrement function is the same for all test takers who have the same switching point. It is further assumed that, in all classes, the decrement is smaller when PD onset occurs later in the test.

Jin and Wang (2014) assume that, after PD onset, the item parameters are the same as before the switching point ($\tilde{\alpha}_i = \alpha_i$ and $\tilde{\beta}_i = \beta_i$), but that the person variable underlying the aberrant response behavior, $\tilde{\theta}_p$, corresponds to the difference between proficiency and the decrement for the respective switching point, that is, $\tilde{\theta}_p = \theta_p - \kappa_g \cdot (I - \delta_p)$. The multigroup MPDM can be formalized by altering Eq. 1 to

$$\eta_{pi} = \begin{cases} \alpha_i \cdot \theta_p + \beta_i, & \text{if } \delta_p \leq i, \\ \alpha_i \cdot [\theta_p - \kappa_g \cdot (I - \delta_p)] + \beta_i, & \text{if } \delta_p > i. \end{cases} \quad (7)$$

As shown in Eq. 7, the decrement in θ occurring after PD onset is a linear decreasing function across switching points with a group-specific decrement parameter $\kappa_g > 0$ as the slope so that, for the last decline class, the decrement is κ_g.[2] Note that Jin and Wang (2014) modeled θ and δ to be independent of one another. However, we suggest applying Eq. 6 in the same manner as described for the HYBRID for two reasons. First, in this way, it is possible to empirically investigate whether the independence assumption that Jin and Wang (2014) propose holds. Second, we wanted to base our comparisons of the three mixture PD models on similar assumptions regarding proficiency and PD throughout all models. In our multigroup version of the MPDM, the proportions of decline classes were defined in the same way as for the HYBRID, with group-specific shape parameters ω_g (Eqs. 2, 3).

Hence, it can be summarized that the multigroup MPDM shares some similarities with the HYBRID: it makes it possible to examine the group-specific (1) relationship**s** between proficiency and PD (ρ_g-parameters), (2) proportions of examinees showing no PD (proportions π_{Ig}), and (3) distributions of PD onsets δ_p (Eq. 2). However, the MPDM differs from the HYBRID because it assumes (4) that the magnitude of PD depends on δ_p, although the size of PD could be group-specific, and (5) that the responses affected by PD still depend on proficiency.

The present investigation

The aim of the present study was to examine the extent to which the multigroup mixture PD models make it possible to assess PD, and whether the conclusions about the presence and group differences in PD that can be drawn from these models differ. In addition, we examined whether accounting for group differences in PD affected the results of the group comparisons of students' proficiencies, and whether the mixture models envisaged differed in their estimates of group differences in proficiencies. Finally, we investigated whether the mixture PD models provided item parameters estimates that differed from those given by the multigroup 2PLM (Suh et al. 2012). To this end, we drew on a LSA of German students who worked on a mathematics test. We fitted the multigroup versions of the 2PDM, the HYBRID, the MPDM, and a standard multigroup 2PLM. We chose school track as a grouping variable, since school track comparisons are at the core of many large-scale educational programs, and school track has been found to be strongly related to PD, as reflected in item position effects (Nagy et al. 2016).

In Germany, after attending primary school, children are assigned to different school tracks based on their school achievement. There is one academic school track (higher secondary school, *Gymnasium*) and there are several non-academic school tracks, including comprehensive (*Gesamtschule*), intermediate (*Realschule*), and lower secondary schools (*Hauptschule*), though the number of non-academic school tracks can vary (Pietsch and Stubbe 2007). Achievement differences between students at academic and

[2] Although Jin and Wang (2014) also discuss quadratic functions for the decrement function, in their empirical application, they found that a linear function was sufficient.

non-academic school tracks are known to be large (e.g., Prenzel et al. 2013). Furthermore, students at the non-academic tracks appear to take participation in LSAs less seriously (Baumert and Demmrich 2001), possibly giving rise to stronger PD. As such, group differences in PD are likely, and group comparisons of students' proficiencies could be affected by group differences in PD (Mittelhaëuser et al. 2015; Nagy et al. 2016).

Research questions

The analyses reported in this article examined both substantive and methodological research questions. From a substantive point of view, we applied the multigroup mixture PD models in order to examine (Q1) the existence of PD in item responses and track differences therein. Regarding the differences between tracks, we examined (Q2) the proportions of students not affected by PD, (Q3) the distributions of onset points of PD, (Q4) the magnitude of PD effects in each group, and (Q5) the relationships between PD behavior and students' proficiencies.

We expected to obtain the following results: regarding Q1, since the assessment considered was a low-stakes test, we assumed that our data would show some degree of PD and, thus, we expected all mixture PD models to provide a better fit to the data than the 2PLM. Regarding Q2–Q4, in line with research on school track differences regarding item position effects (Nagy et al. 2016) and test-taking motivation (Baumert and Demmrich 2001), we expected the amount of PD to be larger for students attending the non-academic tracks. More specifically, we expected the proportion of students not affected by PD to be higher in the academic track (Q2). We did not feel able to derive detailed expectations about group differences in the onset point of PD (Q3), or in the magnitude of PD effects (Q4) because, to the best to our knowledge, such issues have not been investigated previously. We therefore treated Q3 and Q4 as open research questions to be examined in our application.

So far, only few studies have dealt with the relationship between PD and proficiency (Q5). However, as research on response times indicates that low test-taking effort is associated with low proficiency (Wise and DeMars 2005; Wise et al. 2009), it seemed reasonable to assume that less proficient students would be more likely to show PD. We therefore expected that, within each school track, average proficiency would be lower, the earlier PD occurs.

Our second set of research questions targeted methodological issues. Here, we were interested in (Q6) whether accounting for PD affects the estimates of item parameters, (Q7) whether the multigroup mixture PD models differ in the conclusions that can be drawn from them about the prevalence of PD, and (Q8) the impact that PD has on the results of group comparisons of proficiency.

When PD occurs, the intercept parameters of end-of-test items are likely to be underestimated, whereas their slope parameters are likely to be overestimated when a standard 2PLM is employed (Bolt et al. 2002; Oshima 1994; Suh et al. 2012). Thus, regarding Q6, we expected the estimates of the intercept parameters of end-of-test items in the no-decline class to be higher for the mixture PD models than for the 2PLM, whereas we expected to obtain the opposite result for the item discriminations. We expected this result to hold for all mixture PD models, as Suh et al. (2012) found that most mixture PD models showed quite similar behavior in this respect. Regarding Q7, we did not feel

able to derive detailed expectations because the results provided by the models depend on their representation of PD. Therefore, we treated the question of whether the models presented here converge to similar conclusions as an open research question.

Regarding Q8, we expected that not accounting for PD would lead to an underestimation of the proficiency for the decline classes. Thus, we expected to obtain higher estimates of average proficiency when mixture PD models were employed (Q5a). Since PD was expected to be higher at the non-academic school tracks, we further expected the underestimation of proficiency to be more pronounced in the non-academic than in the academic group. As a consequence, accounting for PD was expected to result in smaller group differences in proficiency between both school tracks. However, the question of whether all models lead to a similar reduction in group differences in proficiency remained open.

Methods

Sample and test design

We considered the mathematics achievement test of the first measurement point of a German large-scale longitudinal educational study, "Aspects of learning background and learning development", which was conducted in the federal state of Hamburg (Behörde für Schule und Berufsbildung 2011). Participation in the assessment was mandatory. The test had no individual consequences for the test takers and, hence, it can be regarded as a low-stakes test. The sample consisted of $N = 12{,}182$ students in the fifth grade at different school tracks, an academic track ($n = 5333$ students) and two non-academic school tracks.

The mathematics assessment was presented as a paper-and-pencil test with a fixed item order, consisting of 30 multiple-choice items, scored as correct or incorrect. Missing item responses caused by omitted and not-reached items were also coded as incorrect. A distinction between incorrect and missing responses was not possible because the original item responses were not made available by the primary investigator. Although this means that PD effects were also influenced by the number of not-reached items, this issue does not appear to be of importance in mixture PD models. Suh et al. (2012) found the 2PDM and HYBRID model to perform equally well regardless of whether not-reached items were coded as incorrect or missing. Furthermore, Jin and Wang (2014) explicitly recommended coding not-reached items as incorrect in their MPDM.

Statistical analyses

We applied the three mixture models (2PDM, HYBRID, MPDM) in the multigroup extensions presented and a multigroup 2PLM to the mathematics achievement test data. The grouping variable was school track and we considered two groups—students attending the academic school track, and students attending a non-academic school track. The models were compared by means of the Bayesian information criterion (BIC; Schwarz 1978) and Akaike's information criterion (AIC; Akaike 1987). Group differences in PD were investigated by comparing the group-specific parameters for each model, and Wald tests were used to test for significant parameter differences between groups. Note that some parameters could not be compared across models; therefore, we investigated whether the patterns found in each model led to similar conclusions regarding PD.

Model specification and parameter estimation

For model identification purposes, the mean and standard deviation of the proficiency variable in the no-decline class at the academic track were constrained to 0 and 1, respectively. The item position of PD onset i_0 for the 2PDM was defined by means of model fit comparisons: the model was repeatedly estimated with switching points $\delta_{i_0} = 10, \ldots, 23$, and the switching point was chosen where the BIC value was lowest. Based on the BIC, the model with a PD onset $i_0 = 18$ showed the best fit to the data and was then used for the further analyses.

The models were estimated by means of marginal maximum likelihood estimation using an expectation–maximization algorithm and using numerical integration with 15 integration points in M*plus* 7.4 (Muthén and Muthén 1998–2012). One problem with maximum likelihood estimation for mixture IRT models is that the solution can converge to a local rather than the global maximum (Finch and French 2012). Therefore, usage of multiple random starting values is recommended to ensure replication of the best likelihood value (Lubke and Muthén 2005). The M*plus* code for the estimated models is given in the additional file to this article.

Results

We first present the goodness-of-fit statistics for the four models (Q1) and show how accounting for PD impacted item parameter estimation (Q6). Next, we present the comparisons between students at the academic track [*academic group* (*aca*)] and students at the non-academic school tracks [*non-academic group* (*nac*)] across models; these results are divided into three subsections. In the first subsection, we examine school-type differences in PD model parameters (Q2–Q4). The next subsection discusses our findings on the relationship between PD and proficiency (Q5) and, in the last subsection, we present the results concerning the impact of PD on proficiency estimates (Q5a, Q8). While presenting the results, we also highlight the similarities and differences between the results provided by the different mixture PD models (Q7).

Model fit and item parameter estimates across models

Model fit

The model fit indices (AIC, BIC) for the different mixture PD models as well as for the 2PLM are presented in Table 1. The MPDM had the best fit to the data, but all mixture PD models were better than the 2PLM. Furthermore, the fit indices of the HYBRID appeared to be closer to the MPDM than to the 2PDM.

Table 1 Model fit

Model	No. of free parameters	LL	AIC	BIC
2PLM	63	− 217,085	434,296	434,763
2PDM	91	− 214,224	428,630	429,304
HYBRID	71	− 213,978	428,098	428,624
MPDM	71	− 213,964	428,070	428,596

LL: log likelihood; AIC (BIC): Akaike's (Bayesian) Information Criterion; 2PLM: multigroup 2PLM; 2PDM: multigroup two-class performance decline model; HYBRID: multigroup HYBRID model; MPDM: multigroup multiclass performance decline model

Item parameter estimates for the no-decline class

Accounting for PD is expected to reduce bias in the parameter estimates (intercepts and slope) of items affected by PD (Suh et al. 2012; Wollack et al. 2003). Since the true item parameters were unknown, we were not able to evaluate whether any of the mixture PD models were able to do this. However, by comparing the item parameters of the no-decline class across models, we were able to investigate whether the estimates of the mixture PD models differed from the 2PLM as well as from one another.

In Fig. 1 the estimated item parameters of the no-decline class (intercept, slope) for all models are depicted. For the intercept parameters, we found differences between the 2PLM and the mixture PD models for the end-of-test items. All of the mixture PD models showed higher estimates for the intercept parameters of the end-of-test items than the 2PLM, which means that items appeared to be less difficult than in the 2PLM. The mixture PD models also differed from one another: while the HYBRID and the MPDM appeared to have rather similar parameter estimates, the intercept estimates of the 2PDM were somewhat lower and closer to the 2PLM.

The slope parameter estimates were quite similar across all of the mixture PD models. Slope estimates differed between the 2PLM and the mixture PD models, not only for the end-of-test items but also for the items at the beginning of the test: while, early in the test, the estimates for the decline models were higher than those for the 2PLM, they were lower at the end of the test. These results indicate that, in the 2PLM, end-of-test

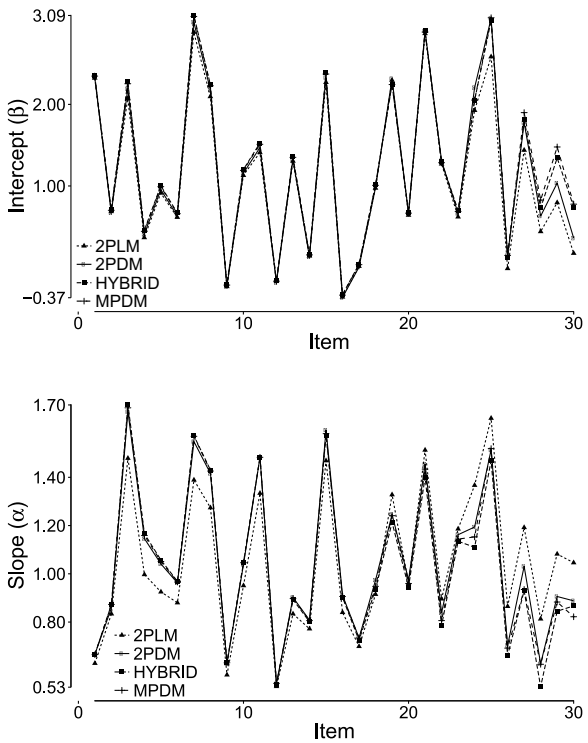

Fig. 1 Estimates of item parameters across models. For the mixture PD models, the displayed item parameters are those of the no-decline class. 2PLM: multi-group 2PL model; 2PDM: multi-group two-class performance decline model; HYBRID: multi-group HYBRID model; MPDM: multi-group multi-class performance decline model

items appeared to be more discriminating, a result also reported by Suh et al. (2012). However, in our application, items positioned early in the test appeared to be less discriminating than in the mixture PD models.

Group differences in performance decline

The main results of all of the group comparisons regarding PD as well as proficiency are displayed in Table 2.

Proportions of no-decline classes and distribution of decline class proportions

The proportions of the no-decline class within each group (i.e. π_{Ig}) are displayed in Table 2, in the section entitled *Distribution of latent classes*. The size of π_{Ig} indicates the number of students who did not show PD during the test. The smaller the π_{Ig}, the smaller the number of students who did not show PD was and, thus, the higher the proportion of students showing PD in the sample was.

Table 2 Parameter estimates of performance decline and proficiency

	nac *Est. (SE)*	aca *Est. (SE)*	Group comparisons Wald's χ^2 (*df*)
Distribution of latent classes			
Proportion of no-decline class (π_I)			
2PDM	0.81 (0.01)	0.91 (0.01)	106.80 (1), $p < 0.001$
HYBRID	0.62 (0.01)	0.82 (0.01)	226.86 (1), $p < 0.001$
MPDM	0.48 (0.03)	0.68 (0.03)	41.60 (1), $p < 0.001$
Shape parameter (ω)			
2PDM			
HYBRID	4.78 (0.15)	7.33 (0.27)	72.27 (1), $p < 0.001$
MPDM	6.47 (0.26)	10.25 (0.47)	70.52 (1), $p < 0.001$
Magnitude of PD			
2PDM	_[a]	_[a]	_[a]
HYBRID: response threshold ($\tilde{\beta}$)	3.76 (0.18)	3.48 (0.31)	0.65 (1), $p = 0.42$
MPDM: decrement (κ)	0.80 (0.04)	0.91 (0.07)	2.85 (1), $p = 0.09$
Proficiency distribution			
No-decline class: mean (μ_I)			
2PDM	− 1.02 (0.03)	0[b]	1637.64 (1), $p < 0.001$
HYBRID	− 1.01 (0.03)	0[b]	1385.61 (1), $p < 0.001$
MPDM	− 1.01 (0.03)	0[b]	1321.48 (1), $p < 0.001$
No-decline class: standard deviation (σ)			
2PDM	0.82 (0.02)	1[b]	133.48 (1), $p < 0.001$
HYBRID	0.82 (0.02)	1[b]	124.16 (1), $p < 0.001$
MPDM	0.82 (0.02)	1[b]	127.80 (1), $p < 0.001$
Decline classes			
2PDM: mean (μ_{i_0})	− 0.95 (0.04)	0.02 (0.08)	167.28 (1), $p < 0.001$
HYBRID: slope (ρ)	− 0.01 (0.004)	− 0.05 (0.01)	17.29 (1), $p < 0.001$
MPDM: slope (ρ)	− 0.01 (0.004)	− 0.03 (0.01)	5.66 (1), $p = 0.02$

Note that parameters cannot be compared across models. See main text for more information

nac: non-academic group; aca: academic group; 2PDM: multigroup two-class performance decline mode; HYBRID: multigroup HYBRID model; MPDM: multigroup multiclass performance decline model. All parameter estimates are group-specific

[a] See Fig. 3. [b] Parameters are fixed

Consistently, in all mixture PD models, π_{Ig} was significantly lower in the non-academic group (Wald's χ^2-test statistics in Table 2), indicating that more students at the non-academic tracks showed PD. However, the no-decline class proportions varied across models. In the 2PDM, the proportions of the no-decline classes were highest: $\pi_{I,aca} = 0.91$ versus $\pi_{I,nac} = 0.81$ (see Table 2, section *Proportion of no-decline class*). As a consequence, the total number of students showing PD was smallest for the 2PDM. In contrast, the proportions of the no-decline class were smallest for the MPDM: here, the no-decline class proportion in the academic group was $\pi_{I,aca} = 0.68$, which means that one-third of the students showed PD. In the non-academic group, the no-decline class proportion was significantly lower, with $\pi_{I,nac} = 0.47$. Thus, one half of the students in the non-academic group showed PD.

While, for the 2PDM, there was only one decline class for each group, for the HYBRID and the MPDM, there were multiple decline classes. As displayed in Eq. 2, the distribution of decline classes was defined by the no-decline class proportion, π_{Ig}, and the shape parameter, ω_g: the higher ω_g, the steeper the increase in decline class proportions across item positions; yet, the higher the π_{Ig}, the smaller the cumulated probability of the decline classes and, hence, the smaller the total number of students showing PD. The cumulated class probabilities across item positions for both models are displayed in Fig. 2. The values for ω_g are displayed in Table 2 (*Shape parameter*).

For the HYBRID, ω_g was significantly higher in the academic group, $\omega_{aca} = 7.33$ versus $\omega_{nac} = 4.78$. Thus, the increase in decline class proportions toward the end of the test was steeper in the academic group (Fig. 2). Similar results were obtained for the MPDM: likewise, ω_g was significantly higher in the academic group ($\omega_{aca} = 10.27$ vs. $\omega_{nac} = 6.47$, Table 2), which means that the increase in class probabilities in the academic group was steeper than in the non-academic group (Fig. 2).

Thus, the HYBRID and the MPDM provided a qualitatively similar distribution of PD classes and group differences within those classes. In both models, the no-decline class was larger in the academic track ($\pi_{I,aca} > \pi_{I,nac}$) and the majority of academic track students showing PD experienced PD onset later in the test ($\omega_{aca} > \omega_{nac}$). However, in the MPDM, class proportions $\pi_{I,g}$ were smaller and shape parameters ω_g were higher in both groups, leading to a steeper increase in cumulated class probabilities for the last five item positions (i.e., for switching points $\delta \geq 24$; Fig. 2). However, in earlier positions,

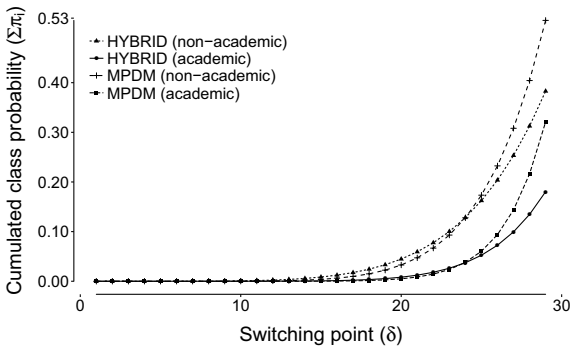

Fig. 2 Cumulated decline class probabilities across item positions for multi-class mixture PD models. HYBRID: multi-group HYBRID model; MPDM: multi-group multi-class performance decline model

the cumulated class probabilities for both models appeared to be almost identical (see Fig. 2). Thus, differences between the decline class proportions of the two models were mainly found for PD onset in the last five items.

Magnitude of performance decline

Group differences in the magnitude of PD were evaluated separately for each mixture PD model according to the respective model parameters assessing the magnitude of PD. In the 2PDM, the group differences found in the item intercept parameters after the switching point (i.e., $\tilde{\beta}_{ig}$, Eq. 4) were regarded as a measure of differential PD effects, while the differences in item intercept parameters between classes affected by PD and no-decline classes (i.e., β_i vs. $\tilde{\beta}_{ig}$, Eq. 4) were regarded as a measure of the magnitude of PD. In Fig. 3, the intercept parameters for the no-decline class as well as for the decline classes for both groups are displayed. The intercept parameter estimates were lower in the decline classes and, thus, the probability of giving a correct response to items after the switching point was lower for students showing PD. Moreover, the intercept parameters decreased across item positions, so that items appeared to become gradually more difficult toward the end of the test in both groups. Furthermore, the results displayed in Fig. 3 do not provide a sound indication of group differences in the onset point of PD as the estimates of the $\tilde{\beta}_{ig}$-parameters decreased immediately after the onset point in both groups. However, the intercept parameters after the switching point were significantly lower in the non-academic group (Wald's $\chi^2(12) = 25.24$, $p = 0.01$), which means that the average size of the PD appeared to be larger in the non-academic school tracks. However, compared to the large difference in intercept parameters between the no-decline and the decline classes, the group differences appeared rather small.

In the HYBRID, group differences in the magnitude of PD were reflected in differences in the response threshold $\tilde{\beta}_g$ (Eq. 5). The response thresholds were $\tilde{\beta}_{nac} = -3.76$ and $\tilde{\beta}_{aca} = -3.48$ for the non-academic and academic group, respectively, which did not differ significantly from one another (see Table 2, section *Magnitude of PD*). After the switching point, the probability of a correct response being provided was 0.02 in the non-academic group and of 0.03 in the academic group of obtaining a correct response after the switching point.

In the MPDM, the probability of obtaining a correct response was not constant after the switching point. Rather, in each decline class, this probability decreased, depending

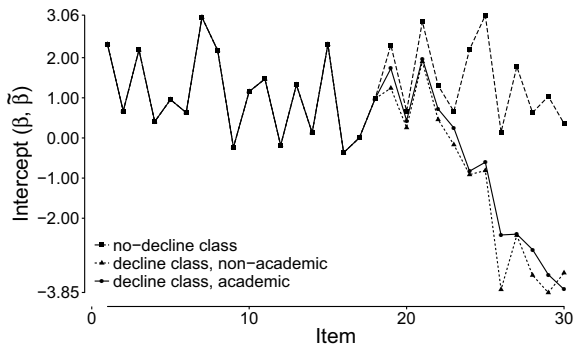

Fig. 3 Intercept parameters of no-decline and decline classes (multigroup two-class performance decline model). Item parameters of the no-decline classes are invariant between groups

on the PD onset point, with group-specific decrement parameters κ_g (Eq. 7). However, κ_g did not differ significantly between groups ($\kappa_{nac} = 0.80$ vs. $\kappa_{aca} = 0.91$, see Table 2, section *Magnitude of PD*), indicating that the magnitude of PD depended only on the PD onset point, and not on the school track the students attended.

Hence, it can be summarized that, of the mixture PD models, the HYBRID and the MPDM provided no indication that the magnitude of PD differs between school tracks. In contrast, the 2PDM revealed statistically significant group differences in the magnitude of PD, with stronger declines in non-academic track students. However, from a substantive point of view, the difference was rather small.

Relationship between performance decline and proficiency

In the 2PDM for both groups, the mean of θ appeared to be somewhat higher in the decline class. In the non-academic group, the mean proficiency was $\mu_{\theta,I,\text{nac}} = -1.02$ in the no-decline class and $\mu_{\theta,i_0,\text{nac}} = -0.95$ in the decline class, but the difference was not statistically significant at the $p \leq 0.05$ level (Table 2, section *No-decline class: mean*). In the academic group, the mean and standard deviation in the no-decline class were fixed at 0 and 1, respectively. The mean proficiency in the decline class was $\mu_{\theta,i_0,\text{aca}} = 0.03$ which did not differ significantly from the no-decline class (Table 2).

For the HYBRID and the MPDM, the mean θ followed a linear function over δ with group-specific slopes ρ_g (Eq. 6). For the HYBRID, in both groups, ρ_g was negative. This indicates that the mean of θ was lower in the decline classes than in the no-decline class and that it further decreased when PD onset occurred in earlier item positions. Furthermore, ρ_g was significantly lower in the academic group ($\rho_{aca} = -0.05$ vs. $\rho_{nac} = -0.01$, Table 2, section *Decline classes*), indicating a stronger negative association between θ and PD in the academic group. More specifically, the value of the ρ_{aca}-parameter of -0.05 means that students who had a PD onset at item 25 (i.e., 5 items before the end of the test) were expected to score 0.25 units lower on the proficiency variable. As the standard deviation of θ was fixed to one in the no-decline class in the academic track, this result directly reflects a standardized effect size.

The results for the MPDM were quite similar to those for the HYBRID. In both groups, ρ_g was also negative and, likewise, the slope in the academic group was significantly lower ($\rho_{aca} = -0.03$ vs. $\rho_{nac} = -0.01$, Table 2). Thus, the average proficiency was lower in the decline classes and the decrease in proficiency when PD onset occurred earlier in the test was steeper in the academic group.

Hence, the HYBRID and the MPDM converged in their conclusion about the relationships between students' proficiencies and their PD onset points, as both models indicated that the earlier the students' PD onset occurred, the lower their proficiency was, and, in both models, this relationship was more exaggerated in academic track students. In contrast, the 2PDM provided no indication of a relationship between proficiency and PD.

Impact of accounting for performance decline on proficiency estimation

Comparisons of individual proficiency scores

In order to illustrate how accounting for PD affects the proficiency distribution, we compared the expected a posteriori (EAP) scores for each model between groups and PD classifications [i.e., students not showing PD (*no-PD class*) vs. students showing PD (*PD*

class)].[3] In order to be able to compare score estimates, we transformed the EAP scores obtained by the PD models to the metric of the 2PLM by means of linear equating (e.g., Livingston 2014).[4]

In Fig. 4, the equated EAP scores for each mixture PD model are plotted against the EAP scores estimated by the 2PLM, displayed separately for school tracks and PD classification. Across all mixture PD models and across both groups, for the no-PD classes, the EAP score estimates of the mixture PD models appeared to be almost identical to those of the 2PLM. However, for students belonging to the PD classes, the EAP scores appeared to be higher when estimated by one of the mixture PD models than those estimated by the 2PLM model. Assuming that scores estimated by the mixture PD models are more accurate reflections of proficiency, the EAP scores estimated by the 2PLM appeared to underestimate the proficiency of test takers showing PD. Looking at the score estimates for the PD classes across models, the correspondence between the EAP scores of the 2PLM and the 2PDM was very high because, in the 2PDM, only one decline class was estimated for each group. For both the HYBRID and the MPDM, the difference in EAP scores between the mixture PD model and the 2PLM depended on the switching point and, thus, the correspondence between scores was lower (Fig. 4). Interestingly, the EAP scores derived from the HYBRID and the MPDM showed very high agreement, although they were based on models that differed in the specification of PD.

Group differences in proficiency

We also investigated whether accounting for PD would affect the estimation of group differences in proficiency. In order to compare the magnitude of group differences as estimated by the four models, we calculated effect sizes between the group means of proficiency. For the mixture PD models, the group means were calculated from the marginal distributions of θ across latent classes.[5] The group means and standard deviations of θ and the effect sizes (d) for all models are displayed in Table 3.

Across all models, the average proficiency in the non-academic group was lower than in the academic group, as expected. In all of the mixture PD models, the effect sizes of the group difference appeared to be very similar to one another (about $d = 1.09$) and smaller than for the 2PLM ($d = 1.17$; Table 3). Thus, groups differed to a smaller degree when mixture PD models were employed, regardless of the specification of PD. However, in the present application, proficiency differences between groups were about one standard deviation for all models, which means that the differences in the results provided by the 2PLM and the mixture PD models are rather small when considered on a relative scale.

[3] For PD classification purposes, the probability of the no-decline class was dummy coded so that students were classified as belonging to the *no-PD class* if their probability of belonging to the no-decline class was larger than 0.5. Otherwise, they were classified as belonging to the combined *PD class*.

[4] In order to obtain equating constants, the subsample of students classified to the no-PD class across all mixture PD models ($n = 8{,}003$) was used. Equating parameters were then applied to all students.

[5] The resulting distribution is a normal mixture distribution with a group-specific mean $\mu_g = \sum_{i=1}^{I} \pi_{ig} \cdot \mu_{ig}$ and a standard deviation as $\sigma_g = \sqrt{\sum_{i=1}^{I} \left[\pi_{ig} \cdot \left(\mu_{ig}^2 + \sigma_g^2\right)\right] - \mu_g^2}$ with $i = 1, \ldots, I$.

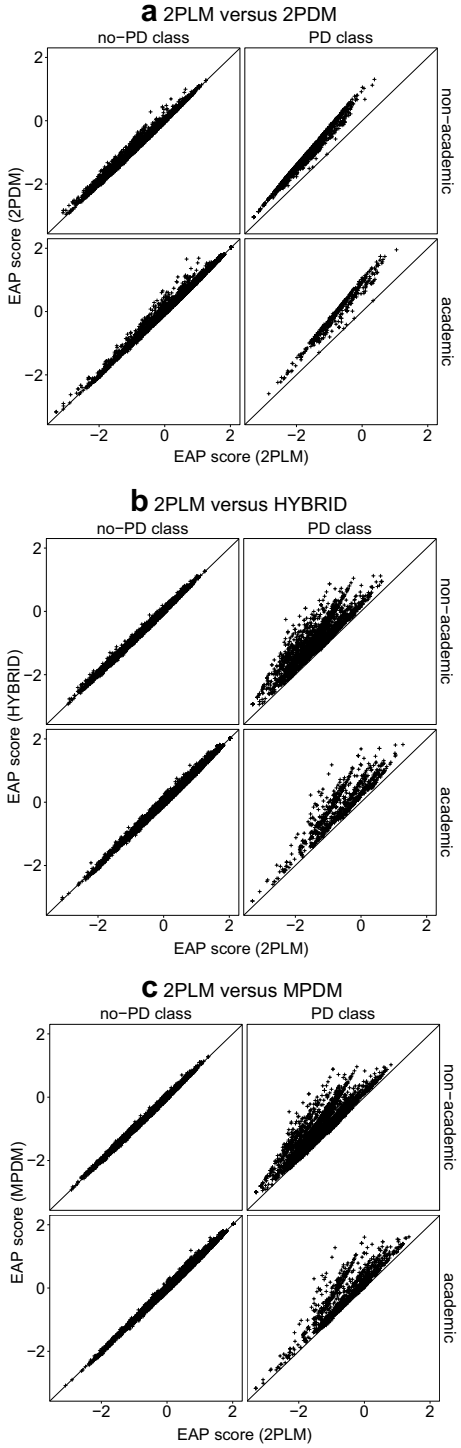

Fig. 4 EAP scores estimated by the 2PLM versus mixture PD models by group and PD classification. x-axis: EAP scores estimated by the multigroup 2PLM; y-axis: equated EAP scores estimated by the multigroup two-class performance decline model (2PDM, **a**), by the multigroup HYBRID model (**b**), and by the multigroup multiclass performance decline model (MPDM, **c**)

Table 3 Group differences in mean proficiency

Model	μ_{nac}	σ_{nac}	μ_{aca}	σ_{aca}	*d (SE)*
2PLM	− 1.07	0.82	0[a]	1[a]	1.17 (0.020)
2PDM	− 1.03	0.82	− 0.04	1.01	1.08 (0.020)
HYBRID	− 1.00	0.82	0.00	1.00	1.10 (0.020)
MPDM	− 1.03	0.82	− 0.03	1.00	1.09 (0.020)

nac: non-academic group; aca: academic group; 2PLM: multigroup 2PLM; 2PDM: multigroup two-class performance decline model; HYBRID: multigroup HYBRID model; MPDM: multigroup multiclass performance decline model. μ and σ obtained from the marginal distribution of θ within groups

[a] Parameters are fixed

Summary and discussion

The aim of the present article was to compare the potential of three mixture PD models, the 2PDM (Bolt et al. 2002), the HYBRID (Yamamoto 1995), and the MPDM (Jin and Wang 2014), for assessing PD in proficiency tests administered in LSAs. The models were extended to accommodate multiple groups, thereby making it possible to examine and test group differences in PD, and to adjust group differences in proficiencies for PD. In order to examine the similarities and discrepancies in the results and conclusions provided by the multigroup mixture PD models, the models were applied to a mathematics tests administered in a German LSA. The results indicate that the models provided similar conclusions about key aspects of PD, but differed with respect to some results. However, the main results gathered by the mixture PD models were in line with results obtained in other settings (e.g., Nagy et al. 2016).

The mixture PD models consistently indicated that the mathematics test was affected by PD, as all of the mixture PD models fitted the data better than a standard 2PLM. Thus, it seems reasonable to assume that, for a subsample of test takers, the responses obtained for end-of-test items reflected PD. In addition, in all of the mixture PD models, the decline class proportions were higher in the non-academic group, but the estimates of the magnitude of PD, such as the intercept parameters of items affected by PD (2PDM), the response thresholds (HYBRID), and the decrement parameters (MPDM) appeared to be similar between groups. Thus, all models indicated that groups differed mainly in the proportions of students showing PD and not in the magnitude of PD. Furthermore, in all models, group differences in the proportions of students showing PD affected the results of the group differences in proficiency. Interestingly, all mixture PD models adjusted the group differences provided by the 2PLM to a similar extent. However, the reductions in the effect sizes were not large when considered on a relative scale; this was due to the very strong track differences in proficiency. Nevertheless, in other settings, where group differences in proficiencies are smaller, the adjustments provided by the mixture PD models might lead to qualitatively different conclusions.

Additionally, all mixture PD models provided relatively similar estimates of item parameters that were different from the parameters estimated by the standard 2PLM. As expected, end-of-test items, that is, those items that were most strongly affected by PD, were estimated to be less difficult than in the 2PLM, and the slope parameters belonging to these items were estimated to be lower by the mixture PD models as compared to the 2PLM. These results are in line with the simulation study of Suh et al. (2012). However,

in contrast to Suh et al. (2012), we found that the mixture PD models provided higher slope parameters for items positioned at the beginning of the test. One reasonable explanation for this result is that individual differences in PD onset points not only increased the dependencies between items at the end of the test, but also reduced the relationships between items at the beginning of test. Hence, when PD was not accounted for, the slope parameters of the 2PLM were estimated to account for the strong dependencies between end-of-test items (i.e., higher slopes for items affected by PD) and the weaker associations between items located at the beginning and the end of the test (i.e., lower slopes for items not affected by PD). Of course, more research is clearly needed to examine the plausibility of our explanation.

Despite the similarities, the three mixture PD models differed from each other in some respects. All of the mixture PD models differed in their estimation of the number of students showing PD: the decline class proportions were smallest in the 2PDM and largest in the MPDM. However, the differences between the MPDM and the HYBRID were rather subtle because the distribution of PD onset points differed only with respect to the last five items. In addition, the mixture PD models provided different conclusions about the relationships between PD onset points and proficiency. In the 2PDM, there were no significant proficiency differences between decline and no-decline classes, neither for students at the academic nor for students at the non-academic school tracks. For both the HYBRID and the MPDM, the mean proficiency was lower in the decline classes and it decreased when PD onset occurred nearer the beginning of the test. Furthermore, in both cases, this decrease was stronger in the academic group. Similarly, the mixture PD models differed in the adjustments of proficiency scores relative to the 2PLM. As we have shown for the EAP scores, for students showing PD, the individual proficiency scores were estimated to be higher in the mixture PD models than in the 2PLM, but the EAP scores provided by the 2PDM were closer to those estimated by the 2PLM. Here, the HYBRID and the MPDM provided EAP scores that were quite similar to each other.

Conclusions and future directions

Taken together, the HYBRID and the MPDM that both consist of many latent classes performed rather similar in many respects, although they differ in their representation of PD. Hence, the number of latent classes combined with the assumptions about their distribution appears to be the main divider between the mixture PD models envisaged in this article. The mixture PD models are based on different assumptions on how PD affects test-taking behavior. In the 2PDM, it is assumed that the switching point is identical for all test takers showing PD within a group. In contrast, in both the HYBRID and the MPDM, multiple switching points are considered, but the models differ in their assumptions about PD. In the HYBRID, it is assumed that the probability of a correct response occurring after the switching point no longer depends on proficiency, whereas, in the MPDM, it is assumed that responses given after the switching point still depend on proficiency. This assumption is also embedded in the 2PDM. However, the question of whether the specification of PD provided by different mixture models are of less importance than the assumptions about the number of PD classes remains open, as suggested by the results provided in this article. Therefore, further research on this issue is called for.

In further applications, the suitability of the proposed model restrictions should be analyzed thoroughly and adapted if necessary. For the HYBRID and the MPDM, we assumed that there was a linear relationship between proficiency and switching points. For some applications, this restriction might be too strict, for example, the decrease in proficiency could be stronger for a switching point in earlier item positions and could diminish when PD affects only the very last items. Moreover, the distribution of decline classes could be modeled by functions other than the one proposed in our study. The issue of model restrictions warrants consideration in subsequent research.

In our study, the HYBRID and the MPDM lead to similar results. Thus, both models appear to be comparably well suited to explore PD in educational LSAs. However, more research is needed on the similarities between these two models.

Finally, the variations of PD effects across domains should be explored. Research on item position effects shows that the strength of the effects varies with respect to the domain being studied (Debeer and Janssen 2013; Nagy et al. 2016). Comparing PD in several domains for the same population could also be useful in determining whether PD can be regarded as an overarching person characteristic or, rather, as a test-specific phenomenon. Mixture PD models with multiple decline classes provide estimates of the switching point and its covariation with proficiency, and this allows for fine-grained analyses of PD. In this sense, mixture PD models can be used to study how PD interacts with proficiency and other variables to provide a better understanding of the mechanisms behind PD.

Additional file

Additional file 1. The Mplus code to analyze the data.

Authors' contributions
OL and GN provided the initial idea for the study. AR and GN provided the initial idea for the statistical analyses. MKL conducted major parts of the statistical analyses. GN and AR conducted minor parts of the statistical analyses. MKL and GN conducted the evaluation of the results. MKL reviewed the literature and wrote the manuscript. AR, OL, OK, and GN read and revised the manuscript. All authors read and approved the final manuscript.

Author details
[1] Leibniz Institute for Science and Mathematics Education, Kiel, Germany. [2] Centre for International Student Assessment, Munich, Germany.

Acknowledgements
The authors would like to thank Gráinne Newcombe for editorial assistance with this article.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
This paper uses data from the longitudinal study "Aspects of learning background and learning development". The data set was generated by the Free and Hanseatic City of Hamburg through the Ministry of Schools and Vocational Training between 1995 and 2005 and have been provided to the MILES scientific consortium (Methodological Issues in Longitudinal Educational Studies) for a limited period with the aim of conducting in-depth examinations of scientific questions. MILES is coordinated by the Leibniz Institute for Science and Mathematics Education.

The *Mplus* code used to analyze our data is provided in Additional file 1.

Funding
There is no funding for this study.

Publisher's Note

Received: 18 April 2017 Accepted: 13 October 2017
Published online: 01 November 2017

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317–332. doi:10.1007/BF02294359.
Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16,* 441–462. doi:10.1007/BF03173192.
Behörde für Schule und Berufsbildung. (2011). *LAU—Aspekte der Lernausgangslage und der Lernentwicklung: Klassenstufen 5, 7 und 9 [LAU—Aspects of learning background and learning development: Grades 5, 7 and 9]*. Münster: Waxmann.
Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–549). Reading: Addison-Wesley.
Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39,* 331–348. doi:10.1111/j.1745-3984.2002.tb01146.x.
Boughton, K. A., & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution rasch models: Extensions and applications* (pp. 147–156). New York: Springer.
Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika, 73,* 209–230. doi:10.1007/s11336-007-9045-9.
Davey, T., & Lee, Y.-H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised general test* (Report No. GREB-08-01). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2011.tb02262.x.
De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement, 35,* 583–603. doi:10.1177/0146621611428446.
De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.
Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39,* 502–523. doi:10.3102/1076998614558485.
Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50,* 164–185. doi:10.1111/jedm.12009.
DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13,* 55–77. doi:10.1207/s15324818ame1301_3.
DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, *8*, 69–82. Retrieved from http://www.rpajournal.com/.
Denis, P. L., & Gilbert, F. (2012). The effect of time constraints and personality facets on general cognitive ability (GCA) assessment. *Personality and Individual Differences, 52,* 541–545. doi:10.1016/j.paid.2011.11.024.
Eklöf, H. (2010). Skill and will: test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17,* 345–356. doi:10.1080/0969594X.2010.516569.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum.
Finch, W. H. & French, B. F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of Maximum Likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods*, 11, 167–178. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol11/iss1/14.
Finn, B. (2015). *Measuring motivation in low-stakes assessments* (Report No. RR-15-19). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12067.
Hailey, E., Callahan, C. M., Azano, A., & Moon, T. R. (2012). An evaluation of test speededness in an assessment for third-grade gifted students. *Journal of Advanced Academics, 23,* 292–304. doi:10.1177/1932202X12462575.
Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage.
Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling, 54,* 418–431.
Jin, K.-Y., & Wang, W.-C. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement, 51,* 178–200. doi:10.1111/jedm.12041.
Livingston, S. A. (2014). *Equating test scores (without IRT)* (2nd ed.). Princeton: Educational Testing Service.
Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10,* 21–39. doi:10.1037/1082-989X.10.1.21.
Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55,* 195–245. doi:10.1007/BF02295283.
Mittelhaëuser, M.-A., Béguin, A. A., & Sijtsma, K. (2013). Modeling differences in test-taking motivation: Exploring the usefulness of the mixture Rasch model and person-fit statistics. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 357–370). New York: Springer.
Mittelhaëuser, M.-A., Béguin, A. A., & Sijtsma, K. (2015). The effect of differential motivation on IRT linking. *Journal of Educational Measurement, 52,* 339–358. doi:10.1111/jedm.12080.
Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
Nagy, G., Lüdtke, O., & Köller, O. (2016). Modeling test context effects in longitudinal achievement data: Examining position effects in the longitudinal German PISA 2012 assessment. *Psychological Test and Assessment Modeling, 58,* 641–670.
Oshima, T. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31,* 200–219. doi:10.1111/j.1745-3984.1994.tb00443.x.

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-scale Assessments in Education*. doi:10.1186/s40536-014-0005-4.
Pietsch, M., & Stubbe, T. C. (2007). Inequality in the transition from primary to secondary school: School choices and educational disparities in Germany. *European Educational Research Journal, 6,* 424–445. doi:10.2304/eerj.2007.6.4.424.
Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (Eds.). (2013). *PISA 2012. Fortschritte und Herausforderungen in Deutschland [PISA 2012. Advances and challenges in Germany]*. Münster: Waxmann.
Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271–282. doi:10.1177/014662169001400305.
Schnipke, D. L., & Pashley, P. J. (1997, March). *Assessing subgroup differences in item response times*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
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,* 213–232. doi:10.1111/j.1745-3984.1997.tb00516.x.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.
Suh, Y., Cho, S.-J., & Wollack, J. A. (2012). A comparison of item calibration procedure in the presence of test speededness. *Journal of Educational Measurement, 49,* 285–311. doi:10.1111/j.1745-3984.2012.00176.x.
van Barneveld, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement, 31,* 31–46. doi:10.1177/0146621606286206.
von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 99–115). New York: Springer.
Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1–17. doi:10.1207/s15326977ea1001_1.
Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163–183. doi:10.1207/s15324818ame1802_2.
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,* 185–205. doi:10.1080/08957340902754650.
Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement, 40,* 307–330. doi:10.1111/j.1745-3984.2003.tb01149.x.
Wu, M. (2010). Measurement, sampling, and equating error in large-scale assessments. *Educational Measurement: Issues and Practice, 29*(4), 15–27. doi:10.1111/j.1745-3992.2010.00190.x.
Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (Report No. TR-95-2). Princeton, NJ: Educational Testing Service.
Yamamoto, K., & Everson, H. (1995). *Modeling the mixture of IRT and pattern responses by a modified hybrid model* (Report No. RR-95-16). Princeton, NJ: Educational Testing Service.
Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the hybrid model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). New York: Waxmann.