Psychological Test and Assessment Modeling, Volume 58, 2016 (4), 641-670

Modeling test context effects in longitudinal achievement data: Examining position effects in the longitudinal German PISA 2012 assessment

Gabriel Nagy[1], *Oliver Lüdtke*[2] & *Olaf Köller*[2]

Abstract

Position effects (PE) in school achievement tests are a specific kind of test context effects (TCEs) that refer to the phenomenon of items becoming more difficult, the later they are positioned in a test. Up until today, PEs have been investigated mainly in cross-sectional settings; this means that little is known about how the size of PEs changes when retesting students. In the present article, we investigate TCEs in the longitudinal extension of the PISA 2012 assessment in Germany. To this end, we propose an extension of the two-dimensional one-parameter item response model, with one dimension per measurement occasion, that includes the effects of booklets (i.e., test forms) on item clusters (i.e., item bundles) that are allowed to vary between assessment occasions and groups (school types). Results indicate that the TCEs uncovered in all domains tested (mathematics, science, and reading) are closely in line with PEs, with reading being most strongly affected, and mathematics being least affected. The size of PEs increased in the second assessment, although the domains were differently affected. This pattern of effects was more pronounced in nonacademic school types. Finally, estimates of average achievement gains appeared to be underestimated by IRT models that neglected TCEs, with differences being largest in domains most strongly affected by PEs (i.e., science and reading).

Keywords: PISA, Item Response Theory, Test Context Effects, Position Effects, Achievement Growth

[1] *Correspondence concerning this article should be addressed to:* Prof. Dr. Gabriel Nagy,Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany; email: nagy@ipn.uni-kiel.de

[2] Leibniz Institute for Science and Mathematics Education, Kiel, Germany

During the last decades, longitudinal studies have become increasingly popular in the psychological and educational sciences and have been adopted in many recent large-scale studies of student achievement, such as NEAP, NEPS, and PISA (Ramm et al., 2006). Such studies place high demands on the psychometric quality of the test, which is repeatedly assessed in order to derive achievement scores that can be compared across time. However, the IRT models routinely employed in large-scale assessments assume that the probability of observing a correct response depends only on the item characteristics (i.e., item parameters) and the students' proficiencies. As a consequence, the influences of the context in which items are presented to the students are neglected, although Brennan (1992), for example, listed a number of contextual characteristics that are likely to affect students' test scores. Well-known examples of test context effects (TCE) include position effects (PE; Leary & Dorans, 1985) and effects of domain orders (DOE; Harris, 1991). PEs refer to the phenomenon of items becoming more difficult, the later they are presented in a test. DOEs apply to tests consisting of items from different domains, such as mathematics, science, and reading. DOEs manifest themselves in changes in item responses as reactions to the sequence of domains that precedes a specific item or a section of the test. Other examples of TCEs are effects of difficulty, caused by the sequencing of items or sections, such as easy-to-hard versus hard-to-easy sequences, effects of testing time, and effects of the ordering of response options, among others. A complete list of all possible kinds of TCEs is hard or even impossible to derive. Nevertheless, large parts of current research agree that PEs are the most prevalent types of TECs and affect almost all school achievement tests (Leary & Dorans, 1985).

TCEs can be regarded as a threat to the validity of inferences about changes in proficiency levels for two reasons. First, in many longitudinal studies, the booklet design is changed across assessments, so that the test scores derived on the different measurement occasions are differently impacted by TCEs. In this scenario, changes in the test design are the sole reason for biased change estimates. As a consequence, group differences in proficiency gains should not be sensitive to TCEs because all individuals are equally affected by the changes in the assessment design. Second, TCEs can be conceived as individuals' reactions to the features of the test form provided (e.g., Debeer, Buchholz, Hartig, & Janssen, 2014), so that the strengths of reactions could differ across time even when the test design remains unchanged. In this scenario, the reason for changes in TCEs is located on the person side, so that group differences in proficiency gains are biased when the size and/or pattern of TCEs changes across time in a group-specific way. Of course, in real applications, changes in TCEs could be due to both reasons (i.e., changes in the assessment design, and changes in individuals' reactions to the test).

In the present article, we investigate TCEs in the German longitudinal extension of the PISA 2012 assessment. This study involved a second assessment of a subsample of the 9th graders who participated in the PISA 2012 study and who were retested when they were in the 10th grade. We propose an IRT model that makes it possible to assess TCEs operating on the level of item clusters, which are the main building blocks of test designs employed in large-scale assessments of student achievement, such as PISA. To this end, we specify booklet effects (i.e., effects of test forms) that are assumed to vary between item clusters. The model furthermore makes it possible to specify these effects to vary between measurement occasions, and between student groups (e.g., school types). These analyses shed light on many highly relevant questions. First, we examined whether the identified patterns of TCEs are in line with the pattern of effects expected if PEs are present. Second, we investigated whether the strengths of TCEs, and possibly PEs, differs between academic and nonacademic school types. Third, we inspected whether the pattern of TCEs indicates changes in the strengths of PEs across time, possibly interacting with school type. Fourth, we examined the consequences of ignoring TCEs (and hence PEs) when examining cross-sectional and longitudinal differences between school types.

Test context effects in school achievement tests

Individuals' reactions to a test can be simultaneously affected by different kinds of TCEs (Brennan, 1992; Leary & Dorans, 1985). For example, research on the effects of section and/or item scrambling (e.g., Harris, 1991; Liu & Dorans, 2012) demonstrates that the scrambling of items or sections in different test forms (i.e., booklets) affects the probability of correct responses and might therefore have adverse consequences for the comparability of results across booklets. Scrambling can lead to TCEs of different kinds, because scrambling is likely to change the context in which an item or a section appears in various ways, such as the position in the test, and the order of domains preceding a given item or section, among others. Inferences about specific types of TCEs, such as PEs and DOEs, could be complicated because the contextual characteristics of booklets are often confounded. For example, the positions of any item or section across booklets could be accompanied by a specific domain order preceding each position.

The confounding of different contextual characteristics is quite common in booklet designs employed in large-scale assessments of student achievement. For example, the booklet designs employed in the international PISA assessments consist of item sections (called item clusters), with each item cluster being presented exactly once in each cluster position (see below). A consequence of this design is that each position of an item cluster is necessarily accompanied by only one specific order of a subset of distinct item clusters. This means that each position of an item cluster is accompanied by exactly one order of domains (in the case of tests composed of multiple domains), and/or a specific order of difficulties of item clusters, among other characteristics. Despite these problems, PEs, as one specific kind of TCEs, have received the most attention in the psychometric literature (Leary & Dorans, 1985). Two basic approaches to the examination of PEs can be distinguished. First, item or cluster scores (or item or cluster difficulties) can be compared across booklets. Situations in which differences in scores or difficulties are related to the item clusters' positions in the test suggest that TCEs are in part due to PEs (e.g., Meyers, Miller, & Way, 2009). Second, in recent years, psychometric measurement models that include PEs have been developed (e.g., Debeer & Janssen, 2013). These models assume a functional form of increases in item (cluster) difficulties across positions, whereas a linear function is typically employed. These models implicitly assume that TCEs can be fully attributed to PEs.

Regardless of the approach taken, research suggests PEs to be the rule rather than the exception in school achievement tests (e.g., Meyers, Miller, & Way, 2009). For example, the achievement tests administered in the PISA assessments appear to be prone to PEs, because item difficulties increase, the nearer the items are positioned towards the end of the test (Debeer, Buchholz, Hartig, & Janssen, 2014; Hartig & Buchholz, 2012; Wu, 2010). More recently, PEs have been discussed as indicators of students' test-taking persistence (Debeer, Buchholz, Hartig, & Janssen, 2014), which means that the strengths of PEs could, in principle, vary across groups, time, and even individuals (Debeer & Janssen, 2013). Indeed, some research conducted on the basis of the PISA assessments indicates that the size of PEs varies between countries and schools, such that PEs are stronger in groups with lower average achievement (Debeer, Buchholz, Hartig, & Janssen, 2014; Hartig & Buchholz, 2012). In addition, Qian (2014) provided some evidence that the strength of PEs is related to a variety of individual and group characteristics, including the school type that the students attend.

However, up until today, little is known about whether TCEs in general, or PEs in particular, change across occasions of measurement. DeMars (2007) investigated changes in the rate of rapid guessing behavior in repeated testing and found that, in low-stakes testing situations, the probability of rapid guessing increases in later assessments. Hence, it could be expected that the same could hold true for PEs because both PEs and rapid guessing could be regarded as indicators of a reduced test-taking persistence.

Although PEs, as a specific instance of TCEs, have been extensively investigated, only a few studies investigated the consequences of ignoring TCEs when comparing test scores derived on the basis of the same items administered in different test booklets. Exceptions to this are the previously mentioned studies on item scrambling (Harris, 1991; Liu & Dorans, 2012), as well as studies investigating the impact of TCEs on test equating procedures (e.g., Leary & Dorans, 1985; Meyers, Miller & Way, 2009). These studies came to the conclusion that TCEs harm the comparability of test scores when they are not accounted for in the equating procedure. The solutions offered so far either rely on specific test designs that should counteract the unwarranted effects (Meyers, Miller & Way, 2009), or on specific equating procedures (e.g., Moses, Yang, & Wilson, 2007). An alternative could be to apply the IRT approach suggested by Debeer and Janssen (2013) that involves randomly varying PEs. However, this method does not appear to be well suited for handling more general forms of TCEs that do not follow an a priori specified function of item orders. In this article, we therefore suggest a simple, but quite flexible IRT model that provides estimates for TCEs on the level of item clusters. In addition, the model suggested sets the metric of the latent proficiency variable to be compared across groups and/or time in reference to the clusters presented in the first position, thereby allowing correcting for TCEs when conducting comparisons.

Assessing test context effects in the IRT framework

In recent years, IRT models have been extended to accommodate PEs as one specific kind of TCEs. The models require rotated booklet designs in which the position of items varies between booklets, and students are randomly assigned to booklets. Most IRT models used for assessing PEs can be regarded as specific instances or extensions of the linear logistic test model (LLTM; Fischer, 1973). They assume item difficulties to be linearly related to item positions (Hohensinn et al., 2008). The approaches have been extended to include random effects located on the level of individuals (Debeer & Janssen, 2013) or items (Weirich, Hecht, & Böhme, 2014).

The IRT models for PEs have an intuitive appeal but may not be optimally suited for situations in which other kinds of TCEs could operate alongside PEs. For example, in tests composed of different domains, DOEs could be at work, which could mean that at least two kinds of specific effects might give rise to TCEs. Furthermore, in test designs intended to assess multiple domains, domain orders are typically confounded with the positions of clusters. For example, in the PISA assessment, the clusters are included in only one booklet in a specific position, which means that each specific position is paired with a specific sequence of domains. Such a design does not allow for a formal separation of PEs and DOEs, among other types of TCEs, but rather calls for an estimation of the compound effects of TCEs of different kinds.

In the present article, we propose an IRT model that aims to estimate TCEs instead of PEs. The results provided by the model can be used to examine the presence of PEs by examining the pattern of TCEs. If the estimated TCEs reflect a (nearly) monotone declining pattern across positions that appear to occur in (almost) all booklets, it seems likely that the TCEs are largely due to PEs. The larger the deviations from this idealized pattern, the more likely it is that other kinds of TCEs exert an influence. Regardless of the pattern of results, the model can nevertheless be used to adjust group and time comparisons for the effect of TCEs. We did so by defining the metric of the latent proficiency variable in reference to the first cluster position in the test, so that comparisons across groups and time are defined with respect to the same reference point that is least affected by TCEs.

Specifying test context effects by means of booklet effects on item clusters

We propose to assess TCEs on the level of item clusters that serve as the building blocks of the test design used in PISA (Frey, Hartig, & Rupp, 2009; OECD, 2014), as well as many other large-scale assessments of student achievement. In test designs characteristic of large-scale assessments, items are first grouped into homogenous clusters, each requiring the same time to be processed. Items belong solely to one cluster, all items included in one cluster are taken from the same domain (i.e., mathematics, science or reading), and the ordering of items within clusters is held constant across booklets. Item clusters are organized into test booklets, each consisting of the same number of clusters. However, depending on the philosophy of the specific study, booklets are either made up of clusters taken from the same domain (e.g., only mathematics clusters), or of clusters taken from different domains (e.g., mathematics, reading, and science). In most recent large-scale assessments, the test designs employed are balanced with respect to item cluster positions. Typically, each cluster is presented exactly one time in each position, so that each cluster is included in P booklets in positions p = 1, 2, …, P. Students are randomly assigned to booklets, thereby ensuring that the groups of students working on the different booklets are (randomly) equivalent. As a consequence, there are no systematic differences in the proficiency distributions between student groups working on different booklets.

Given the random assignment of students to booklets, TCEs can be examined for each cluster by comparing the results for a specific cluster across booklets. Because student groups defined by booklets are randomly equivalent, differences in results can be attributed to TCEs. We suggest quantifying TCEs in reference to the booklet in which a specific cluster is included in the position $p = 1$. The first position is a natural starting point because the conditions at the beginning of the testing session can be expected to be (randomly) equivalent for students working on different booklets. Hence, differences in results gathered on the basis of a specific cluster presented in a later position ($p > 1$) could be indicative of TCEs of different kinds, such as PEs. Note, however, that, depending on the peculiarities of the booklet design employed, most designs do not allow different kinds of TCEs to be precisely identified.

Although typical test designs do not allow for a formal separation of TCEs of different kinds, such as PEs and DOEs, the identified patterns of TCEs can be inspected in order to examine whether they provide an indication for specific types of TCEs. As the test designs employed in large-scale assessments are typically balanced for item positions, but not necessarily for other characteristics, it is easiest to check whether the patterns of TCEs indicate that PEs are at work. Regarding PEs as reflections of individuals' persistence when working on items assessing proficiency in a given domain (e.g., Hartig & Buchholz, 2012), a (nearly) monotone decline in average scores in clusters can be expected. Such a pattern indicates that all item clusters exhibit similar patterns of TCEs across positions and appear to be only weakly related to differences in other contextual characteristics between test booklets. Of course, such a result does not formally rule out alternative explanations because each position of a given item cluster is confounded with other characteristics of the booklets (remember that item clusters are typically only presented once in each position). However, as, in practice, the item clusters presented in the same position are surrounded by different configurations of test characteristics, it appears rather unlikely that the different patterns of test contexts lead to almost identical patterns of TCEs that show a close fit to the expected patterns implied by PEs. Hence, we argue that a pattern of score declines across positions within booklets that appears to be similar across booklets suggests that PEs are one of the main driving forces underlying TCEs. However, such a result should definitely not be taken as an indication that the observed TCEs are solely due to PEs.

An IRT model for test context effects

The IRT approach suggested in this article is based on the one-parameter logistic test model (1PL), also known as the Rasch model (Rasch, 1960). We chose the 1PL because the PISA framework is based on this model (OECD, 2014) and we intended to apply the model to the longitudinal PISA 2012 assessment. However, as PISA has recently switched to the 2PL model, and many other testing programs rely on the 2PL model, an extension of our IRT approach to TCEs to the 2PL appears worthwhile. Although we do not pursue such a development in the present article, we note that the ideas presented can also be easily applied to the 2PL framework. We present a specification that includes two measurement occasions (t = 1, 2) that are accommodated by a two-dimensional model (von Davier, Xu & Carstensen, 2011). The model is further specified in a two-group context (g = 1, 2), allowing the size of TCEs to be compared between groups.

Let y_{ijcbtg} be the item response observed for individual i to item j belonging to cluster c included in booklet b assessed on occasion t in group g. In the case of dichotomous item responses, we model the logits of the probability of correct responses as

$$Log\left[\frac{P\left(y_{ijcbtg}=1\right)}{P\left(y_{ijcbtg}=0\right)}\right]=\theta_{it}+\nu_{ict}-\beta_{j}\ , \tag{1}$$

with θ_{it} indicating the latent proficiency variable for individual i on occasion t, and β_j standing for the difficulty of item j. Hence, we assume that the ordering of item difficulties within a cluster is invariant across groups and occasions, but this assumption can be relaxed. The term ν represents a node variable related to the items in cluster c assessed on occasion t, which means that the model consists of as many node variables as there are combinations of clusters c and measurement occasions t. Note that, although ν is indexed by the person subscript i, this does not mean that we allowed the ν-variables to freely vary across individuals. Instead, we considered the node variables to be fully determined by person covariates (see below).

Ordinal items with more than two categories are modeled by extending Equation 1 to the partial credit model (Masters, 1982). The logit of the probability of taking the step from category l to the next highest category m is given by

$$Log\left[\frac{P\left(y_{ijcbtg}=m\right)}{P\left(y_{ijcbtg}=l\right)}\right]=\theta_{it}+\nu_{ict}-\tau_{jm}\ , \tag{2}$$

where τ_{jm} stands for the threshold separating the successive categories l and m in item j.

The node variables ν included in Equations 1 and 2 are specified to depend on booklet indicators d_{ibt}, that indicate whether person i had received booklet b on occasion t (d_{ibt} = 1) or otherwise (d_{ibt} = 0):

$$\nu_{ict}=\sum_{b=1}^{B}\gamma_{cbtg}d_{ibt}\ . \tag{3}$$

The γ-parameters given in Equation 3 quantify the TCEs operating on the level of clusters. In line with the arguments outlined previously, the γ-parmeters of the booklet in which cluster c is presented in the first position are fixed to 0. The same applies to booklet effects on item clusters not included in the corresponding booklets. These constraints imply a clearly defined point for anchoring the θ-variables, so that they are specified in reference to the clusters presented in the first position of each booklet.

Node variables, as used in Equations 1 to 3 in the present article (ν-variables), have long been used in applications of structural equation models. They appear under different labels, such as phantom variables (e.g., Rindskopf, 1984), and serve different purposes depending on the aim of the analysis. Hauser and Goldberger (1971) introduced node variables to impose proportionality constraints on the effects of explanatory variables on multiple outcomes in path analyses. Our use of node variables is close to this approach. Inserting Equation 3 into Equations 1 or 2 shows that, in the model presented, it is assumed that the TCEs represented by the γ-parameters in Equation 3 apply to all items included in one item cluster. This assumption can, in principle, be relaxed by including extra parameters that allow items to be affected differently by TCEs. However, we have not included such an extension in the present article, because such a model does not appear to be compatible with the basic assumptions underlying the Rasch model. Additional parameters would be necessary if our approach to TCEs is to be extended to the 2PL model.

For the proficiency variables, we assume a bivariate normal distribution with mean vector $\boldsymbol{\alpha_g} = \left[\alpha_{1g}, \alpha_{2g}\right]'$, and covariance matrix $\boldsymbol{\Phi_g} = \begin{bmatrix} \Phi_{11g} & \\ \Phi_{12g} & \Phi_{22g} \end{bmatrix}$, whereby the subscript g indicates that these entities are allowed to differ between groups. In order to keep the model identified, one mean parameter α_{tg} needs to be fixed (e.g., $\alpha_{11} = 0$). Because the proficiency variables are anchored in reference to the clusters presented in the first booklet positions, mean differences between groups, as well as mean differences across time are given relative to the first cluster positions. Hence, if one agrees that responses given in the first cluster position are relatively free of TCEs, the differences derived in this way are also barely affected by TCEs.

The model proposed can be estimated by means of conventional software packages, but its implementation could be complicated by the fact that the booklet indicators used in Equation 3 are linearly dependent. Simply stated, if we know the individuals' values on the first $B - 1$ booklet indicators on assessment occasion t, $d_{(b=1)t}$ to $d_{(b=B-1)t}$, we can deduce their values on the last booklet indicator $d_{(b=B)t}$. Such a situation does not allow the regression part given in Equation 3 to be estimated as no unique solution exists for all γ-parameters. To overcome this problem, Equation 3 can be reframed by using one booklet per measurement occasion as a reference. For example, in the case of B booklets on occasion t, we might use the last booklet $b = B$ as a reference, meaning that only $B - 1$ booklet indicators are used on this occasion. Suppose that cluster c is included in booklet $b = x$ in the first position: in order to fully capture booklet effects on clusters, Equation 3 can be reformulated as

$$\nu_{ict} = \gamma_{c(b=B)tg} + \gamma^{*}_{c(b=1)tg} d_{i(b=1)t} + \ldots - \gamma_{c(b=x)tg} d_{i(b=x)t} + \ldots + \gamma^{*}_{c(b=B-1)tg} d_{i(b=B-1)t} \,. \tag{4}$$

The trick applied in Equation 4 is that the TCE associated with booklet $b = B$ is absorbed in a constant term not included in Equation 3, and the effect associated with booklet $b = x$ is constrained to be the negative value of the constant. In this situation, the effects associated with the remaining booklet indicators no longer correspond to the TCEs given in Equation 3 and are therefore flagged by an asterisk. These parameters are related to the γ-parameters in Equation 3 by $\gamma^{*}_{cbtg} = \gamma_{cbtg} - \gamma_{c(b=B)tg}$, such that $\gamma_{cbtg} = \gamma_{c(b=B)tg} + \gamma^{*}_{cbtg}$.

Figure 1 gives an example of the parametrization outlined in Equation 4 by means of a path diagram. The figure recurs on a hypothetical example consisting of two measurement occasions on which three booklets were administered, each composed of two out of three clusters (B1T1: 1,2; B2T1: 3,1; B3T1: 2,3; B1T2= 1,3; B2T2 = 3,2; B3T2: 2,1). To keep the example simple, we assume a situation with a single group so that the group subscripts are omitted. In Figure 1, latent proficiency variables θ_t are represented by white circles, whereas node variables ν_{ct} are depicted as gray, shaded circles. Rectangles represent either items (left blank) or booklet indicators (d_{bt}). (Co-)variance terms are represented by bidirectional arrows. Directed arrows stand for booklet effects on node variables (γ-parameters), and directed arrows starting from a triangle represent latent means (α_t) and constant terms connected with node variables (γ-parameters). In this example, only the node variables ν_{31} and ν_{12} have constant terms attached to them. This is done to capture TCEs associated with booklet indicators d_{31} and d_{32} (i.e., γ_{331} and γ_{132}), which are not included in the model because they served as reference points. Because the TCEs of clusters presented in the first position of a booklet are constrained to be zero, the effects of booklet indicator d_{21} on ν_{31} is constrained to be the negative value of the intercept term γ_{331}. This ensures that the TCE for cluster 3 included in the second booklet, where it is presented in the first position, evaluates to 0 for examinees who received booklet two [i.e., $\gamma_{331} + (-\gamma_{331})d_{21} = \gamma_{331} - \gamma_{331}$]. The same logic applies to

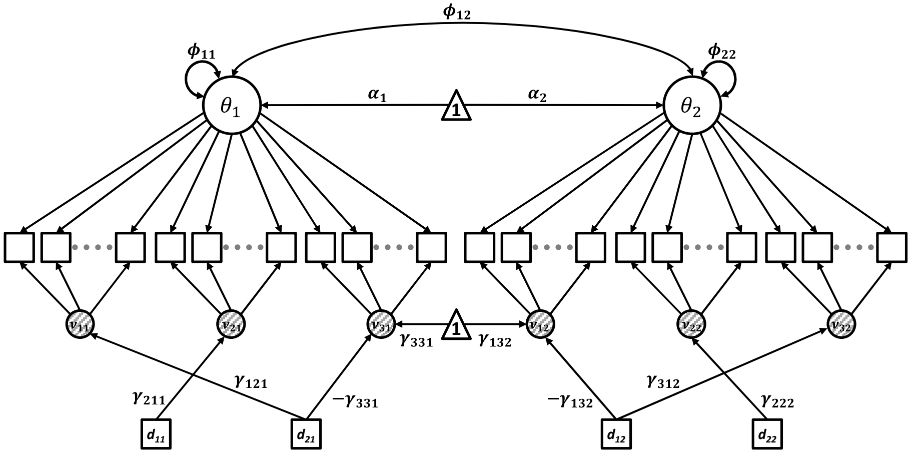

Figure 1:
A hypothetical application of a longitudinal IRT model including test context effects represented as a path diagram. To ease the presentation, parameters are not flagged by group membership. See main text for explanations

cluster 1, assessed on the second occasion. For examinees who worked on booklet one, where the cluster was presented in the first position, the TCE evaluated to 0. The remaining γ-parameters included in Figure 1 directly represent the cluster-specific TCEs associated with booklets in which they appear in the second position. Therefore, in the example, there is no need to derive TCEs from γ^*-parameters.

Software implementation

The IRT model described in this section can be estimated by means of marginal maximum likelihood techniques employing the expectation maximization algorithm. Such estimation routines are implemented in many software packages for latent variables. We employed M*plus* 7.4 (Muthén & Muthén, 1998-2012), but other programs, such as OpenMx (Neale et al., 2016) and EQSIRT (Wu & Benter, 2011), might be used as well. Note that the model requires only as many dimensions of integration as there are latent proficiency variables included in the model. Although the node variables are formulated as (pseudo-) latent variables, they technically only serve the purpose of distributing the TCEs represented by the γ-parameters across all items included in a given combination of clusters and booklets.

We chose the M*plus* program because of pragmatic reasons. For example, M*plus* makes it possible to deduce the standard errors of the more fundamental γ-parameters reflecting TCEs on the basis of the estimated γ^*-parameters by applying the delta method. In addition, the program allows the summary statistics of γ-parameters to be computed alongside their corresponding standard errors. For example, γ-parameters can be averaged across item clusters for each position, and the average effects can be compared across time and/or groups. In the present investigation, we made use of this option to describe the general trends and group differences in position-specific TCEs averaged over clusters.

The present investigation

In the present article, we examine TCEs in the longitudinal extension of the PISA 2012 assessment in Germany. This study comprised a subset of schools and students that were in Grade 9 on the first measurement occasion and in Grade 10 in 2013, therefore the focus of the longitudinal study is on grade-based instead of age-based samples. Because of the assessment design implemented, the lowest track school type (German "Hauptschule"), is no longer part of the target population (as the "Hauptschule" ends with Grade 9). Hence, the student population reflected in the longitudinal sample is characterized by two broad groups attending different kinds of schools. Here, we considered students from the academic track (i.e., the Gymnasium) as a separate group, because these students are known to exhibit test scores that are, on average, well above the levels of other school types. The second group comprises students from the middle track (i.e., the Realschule), as well as from different forms of comprehensive schools that have been reported to exhibit roughly comparable achievement levels. PISA consists of achievement tests in the domains of mathematics, science, and reading. These domains were also assessed on both occasions in the longitudinal extension study.

In the present research TCEs were examined from a cross-sectional and longitudinal perspective. Regarding the cross-sectional perspective, we expected to obtain the following pattern of results: First, in line with previous studies investigating PEs in the PISA instruments, we expected TCEs to occur, whereby the pattern of TCEs was expected to be indicative of PEs on both occasions (i.e., decreases in item cluster scores across positions). Second, based on the results reporting stronger PEs in student groups with lower achievement (Debeer, Buchholz, Hartig, & Janssen, 2014; Hartig & Buchholz, 2012), we expected the TCEs to reflect stronger PEs in the nonacademic group on both occasions (i.e., steeper declines in item cluster scores across positions in the nonacademic group). We explicitly expected this pattern to occur in the domains of reading and science, because these domains have been investigated in previous studies. As group differences in TCEs in mathematics have not yet been examined, we treated this issue as an open research question.

To the best of our knowledge, TCEs in general, and PEs in particular, have never been investigated in longitudinal settings. However, due to initial evidence showing increases in the prevalence of unmotivated test-taking behavior in retest situations (DeMars, 2007), TCEs indicating increases in PEs across time do appear plausible. Such a pattern intuitively makes sense as the PISA assessment appears likely to elicit aversive reactions because it requires students to work for two hours on tests without revealing any explicit rewards. Furthermore, in light of the previous studies reporting stronger PEs in lower achieving student groups, it appeared reasonable to expect that PEs would become stronger in the nonacademic tracks. Therefore, from the longitudinal perspective, we first expected TCEs to indicate stronger PEs (i.e., steeper declines across positions) on the second measurement occasion. However, given the sparse research findings so far, it was not clear to us whether this effect occurs for all domains assessed in PISA. Second, we examined the plausible possibility that the revealed TCEs point towards a pattern of PEs that become more strongly accentuated across time in the nonacademic group compared to the academic track students.

The next set of research questions concerns the consequences of ignoring TCEs for the results of mean comparisons within and across time. Differences in TCEs between groups and assessment occasions are likely to affect the results of mean comparisons. This is especially the case in situations where groups and/or assessment occasions are differentially impacted by TCEs. When TCEs are not controlled for, these effects are absorbed into the estimates of the students' proficiency distributions, thereby affecting the results of mean comparisons. Although it is well known that TCEs counteract the comparability of test scores derived on the basis of different orders of test material (e.g., Liu & Dorans, 2012), their consequences for mean comparisons have – to our knowledge – not been systematically evaluated so far. One aim of the present paper is to examine the consequences of ignoring TCEs when conducting cross-sectional and longitudinal mean comparisons. We approached this problem by comparing the results provided by the model suggested in this article with the results provided by a conventional IRT model that ignores TCEs. We expected TCEs to make a difference, and expected their effect to depend on the comparison undertaken.

In the case of cross-sectional comparisons between school types, we expected group differences in proficiency levels favoring the academic track students to be somewhat reduced when TCEs were controlled for. This is because we expected TCEs to reflect stronger PEs in students from nonacademic schools that are absorbed in the estimates of proficiency levels when TCEs are not controlled for. However, we did not expect TCEs to have a strong impact on the cross-sectional results because the differences in average proficiency levels are known to be large. Hence, from a practical point of view, the reductions in mean differences after controlling for TCEs were expected to be small on a relative scale. In contrast, in the case of longitudinal comparisons, where effect sizes for group differences in change are usually smaller, results are likely to be more sensitive to TCEs. Here, we expected that accounting for TCEs could, theoretically, lead to qualitatively different conclusions about group differences in proficiency growth. Large parts of the empirical literature report rather small group differences in growth rates that are, in principle, more sensitive to measurement artifacts, such as TCEs. However, the impact that the adjustment of TCEs has on the estimated group differences in proficiency gains clearly depends on the size of the group differences in longitudinal changes in TCEs. Whether and to what extent the adjustment of TCEs affects group differences in growth is therefore an open research question that was approached in the present investigation.

Method

Sample and booklet design

The sample consisted of $N = 6359$ students (50.7% females) who were included in the German sample of the international PISA assessment which took place in 2012. The students considered in the present research were a subsample because they were all in Grade 9 in 2012, and were not students attending the German "Hauptschule" which ends after Grade 9. Furthermore, we included only those students into the analyses who attended schools which participated in the second assessment in 2013. As participation in the retest assessment was voluntary, a number of schools did not participate in 2013 and all students from these schools were excluded from the sample. Participation in the retest occasion was also voluntary at the school and student level. Of the initial sample, 67.2% of students ($N = 4271$) participated on the second occasion. Additional analyses revealed that initial achievement was the most important predictor of schools' and students' participation in the second assessment (Heine, Nagy, Meinck, Zühlke, & Mang, 2016). Because we included initial proficiency levels in all analyses, our main results can be expected to be relatively robust against biases due to selective dropout (e.g., Little & Rubin, 2014). Hence, we expected our analyses to estimate TCEs and their changes in the German PISA population as defined in the present article.

Table 1:
Booklet Design Employed in the German Longitudinal Extension of the PISA 2012 Assessment. Items Measuring the Attainment of the German National Educational Standards are Excluded

	Position 1	Position 2	Position 3	Position 4	Students
Grade 9					
Booklet 01	M5	S3	M6	S2	489
Booklet 02	S3	R3	M7	R2	485
Booklet 03	R3	M6	S1	M3	481
Booklet 04	M6	M7	R1	M4	504
Booklet 05	M7	S1	M1	M5	490
Booklet 06	M1	M2	R2	M6	507
Booklet 07	M2	S2	M3	M7	498
Booklet 08	S2	R2	M4	S1	489
Booklet 09	R2	M3	M5	R1	493
Booklet 10	M3	M4	S3	M1	478
Booklet 11	M4	M5	R3	M2	471
Booklet 12	S1	R1	M2	S3	485
Booklet 13	R1	M1	S2	R3	489
Grade 10					
Booklet 01	M2	S2	M3	M7	310
Booklet 02	M4	M5	R3	M2	291
Booklet 03	S2	R3	R2	S3	327
Booklet 06	M7	R2	-----	-----	340
Booklet 07	-----	-----	M2	MR	350
Booklet 08	-----	-----	M7	MR	356
Booklet 09	M2	MR	-----	-----	355
Booklet 10	S3	M4	-----	-----	344
Booklet 11	-----	-----	R3	S2	342
Booklet 12	-----	-----	M4	R3	342
Booklet 14	M4	S3	-----	-----	339
Booklet 15	-----	-----	MR	M4	337
Booklet 16	R2	M7	-----	-----	321
Booklet 17	MR	M7	S3	R2	326
Booklet 18	-----	-----	S2	M2	329

Notes: M = Mathematics, S = Science, R = Reading

Of the 6359 students analyzed here, 45.2% were from an academic track and 54.8% from a nonacademic track. However, as we will explain later, the number of students that provided data for each domain tested in PISA was smaller than 6359 in some cases, so that the sample size differed between the analyses targeting the different domains. More specifically, for mathematics, the full sample size was available, while in science, data from N = 4930 students and, in reading, from N = 4954 students were analyzed. The reason for this reduction is that a number of students did not receive test booklets containing any test items targeted at science or reading (see below).

The analyses recurred on items used in the PISA assessments. Table 1 presents the booklets used in 2012 and 2013. In 2012, the booklet design of the international assessment was used. It consists of 13 booklets, with each booklet comprising 4 clusters. In 2012, mathematics was the core domain, i.e., the number of mathematics clusters (7 clusters) was larger than that of science and reading (3 clusters each). As shown in Table 1, each cluster was presented exactly once in each cluster position, so that cluster positions were balanced across booklets.

In the retest assessment (2013), a different design with 18 booklets was used (Table1). From these, 15 booklets included items used in PISA, and only these booklets were used in the analyses reported in this article. In 2013, some clusters used in 2012 were assessed again, but the test also included new items. With the exception of a cluster taken from another PISA assessment (Item cluster MR in Table 1; Ramm et al., 2006), the newly assessed items were not further considered in this article because these items are based on a different framework (national educational standards) so that their inclusion could distort the homogeneity of the test material used in PISA. Furthermore, the number of PISA clusters was largely reduced relative to the first measurement occasion. As shown in Table 1, the test design used in 2013 was no longer balanced with respect to cluster positions. For example, in reading, only one cluster was included in the first position. Nevertheless, the test constructors intended to keep the average cluster position in the second assessment close to the average positon in the first assessment (2.5th for all clusters). This goal was closely met for mathematics (2.4th position), and science (2.5th position), but reading was, on average, assessed in a later position (2.8th position). The maintenance of a balanced item design and the full assessment of all clusters used in 2012 was not feasible with a manageable number of booklets, because the test design was also used to assess test material intended to test students' attainment of the German national educational standards.

Statistical analyses

Analyses were performed by applying the modeling approach suggested in this article separately to each domain assessed in PISA (i.e., mathematics, science, and reading). These analyses provided estimates of TCEs on the level of clusters (i.e., γ-parmeters), which were first inspected for the presence of PEs. We judged TCEs to be indicative of PEs when the TCEs showed a downward pattern across positions. Whenever TCEs appeared indicative of PEs, TCEs were averaged across clusters separately for each position, to get an estimate of the pattern and size of the PEs operating at the level of different school types and measurement occasions. These estimates were further compared to examine whether (1) PEs were higher in nonacademic school types, (2) PEs tended to become stronger on the second occasion, and (3) the changes in PEs differed between school tracks.

The final set of analyses investigated the consequences of ignoring TCEs for examining mean differences in proficiency distributions. To this end, we compared the results provided by the model suggested in this article with the results provided by a traditional IRT model ignoring TCEs. The reference model chosen was a two-dimensional 1PL assuming cross-time and cross-group invariance of item difficulties (cf., von Davier, Xu, & Carstensen, 2011). This model differs from the alternative IRT models that include TCEs (Equations 1 to 3) in its more rigid invariance assumption concerning the item difficulties. The models including TCEs assume full cross-time and cross-group invariance only for the items included in clusters presented in the first item cluster position in each booklet. Items belonging to any cluster presented in a later position $p > 1$ are allowed to have difficulties that differ from the difficulties of items presented in the first item cluster position by a group and time specific constant.

It might be argued that the differences in the results of the alternative models could in part stem from the fact that the models including TCEs might compensate for some alternative forms of violations of parameter invariances across groups and/or across time. However, such deviations from the ideal of full measurement invariance do not depend on the test booklet, whereas the suggested models for TCEs are only capable of identifying violations of measurement invariance that depend on the test booklets employed. Nonetheless, in order to check for meaningful violations to the cross-group and cross-time assumption of invariance of item difficulties, we conducted additional analyses (not reported in this article) and found no evidence for such deviations (i.e., very close agreement between freely estimated item parameters across groups and across time).

All analyses were conducted by means of the M*plus* 7.4 program (Muthén & Muthén, 1998-2012), employing MML estimation via an accelerated EM algorithm, recurring on standard integration with 15 integration points per dimension. Model fit was assessed by the log-likelihood statistic, the AIC, and the sample-size adjusted BIC (sBIC) indices. The AIC and BIC are information-driven measures of fit that account for the model log likelihood and sample size (sBIC) and include penalty functions for the number of parameters (AIC, and sBIC). Models associated with smaller values of AIC and sBIC are preferable, but the values of these statistics cannot be interpreted in isolation. An M*plus* input file is given in the Appendix of the present article for the science domain. Note that, in this example, there is no need to translate γ^* - to γ-parameters (cf. Equation 4).

Results

Our results are presented in two larger subsections. We first present the results regarding TCEs (i.e., differences between school types and occasions of measurement). In the second subsection, we examine the consequences of including versus ignoring TCEs in examining cross-sectional and longitudinal group differences in achievement.

Examination of test context effects

TCEs and PEs. Table 2 presents information about the model fit of IRT models including TCEs and the alternative IRT models ignoring TCEs. As shown in the table, the models that included TCEs outperformed the baseline models in each domain in each measure of model fit. Estimates of TCEs are displayed in Figure 2. Each dot stands for the estimated γ-parameter in a specific booklet and position. γ-parameters belonging to a common booklet are connected by a dotted line alongside the sequence in which clusters were administered. Several observations can be made about the pattern of TCEs.

First, almost all TCEs were negative and, in the majority of cases, became more extreme, the later a cluster was included in a booklet. This pattern of effects indicates that the TCEs identified are likely to be strongly impacted by PEs. Second, the strengths of PEs (average trend of TCEs across positions) appeared to vary between achievement domains, such that mathematics was least and reading most strongly impacted by PEs. Third, the decreasing trend of performance levels appeared to vary in magnitude between school types. Students from nonacademic schools were more strongly impacted by PEs than students from academic schools. Fourth, PEs in nonacademic schools appeared to become stronger in the second assessment, whereas in the academic schools, PEs seemed to be quite constant across time.

School type differences in TCEs. In order to test the visually identified pattern of effects for statistical significance, γ-parameters referring to clusters presented in the same position were averaged and their standard errors were estimated by applying the delta

Table 2:
Fit of IRT Models Excluding (2-Dim-1PL) and Including (2-Dim-1PL+Booklet) Booklet Effects

	Mathematics	Science	Reading
2-Dim-1PL			
# Parameters	120	65	54
Log Likelihood	124722.2	78181.6	55676.5
AIC	249684.4	156493.1	111460.9
sBIC	250114.0	156709.3	111640.8
2-Dim-1PL+Booklet			
# Parameters	194	95	86
Log Likelihood	124462.7	77979.6	55349.0
AIC	249313.4	156149.1	110869.9
sBIC	250007.9	156465.1	111156.3

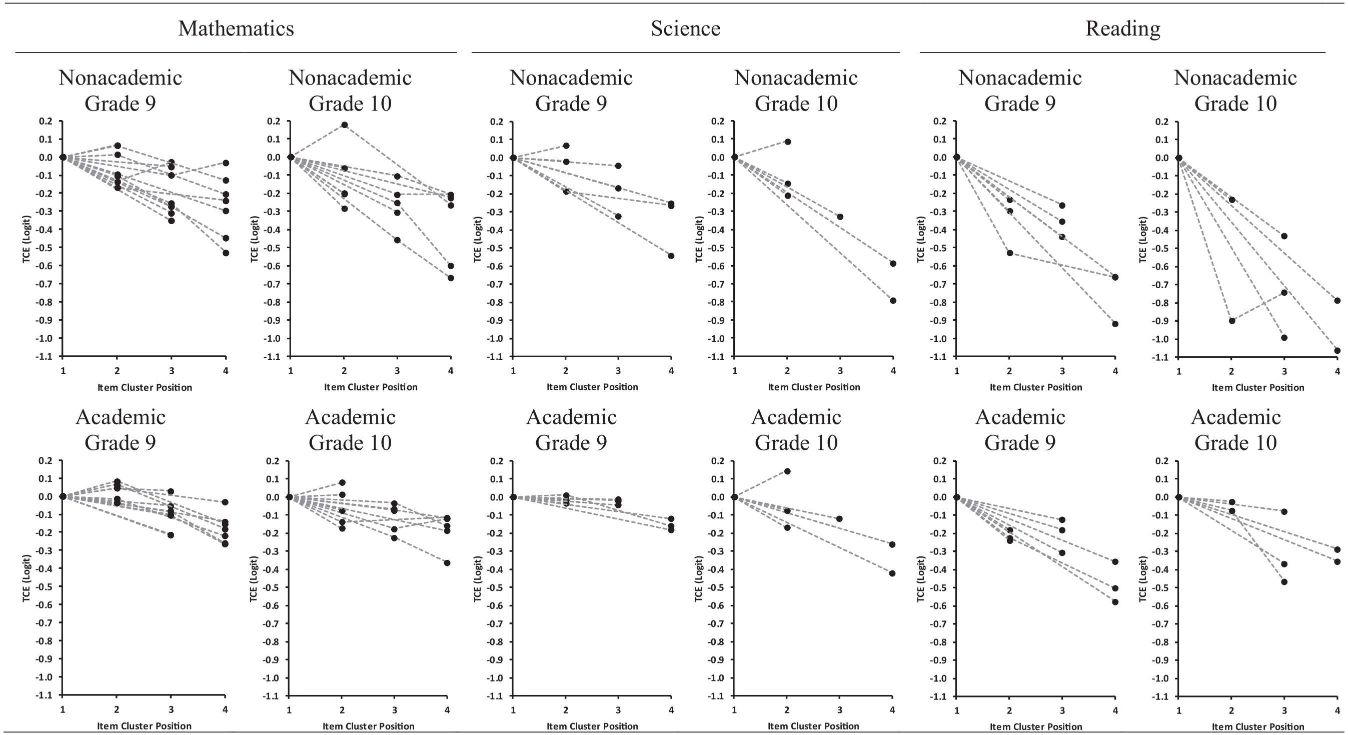

Figure 2:
Text context effects on the level of item clusters presented by item cluster positon (x-axis) for the first (upper panels) and the second measurement occasion (lower panels). Text context effects assessed in one booklet are connected by a dotted line. Effects are presented

method. The corresponding results, along with school type differences in position-specific effects, are summarized in Table 3. As becomes evident, average TCEs tended to exhibit a downward pattern in accordance with PEs. In addition, within measurement occasions, the strengths of effects differed between school types. Effects were generally stronger in nonacademic school types (indicated by negative values of differences). School type differences were weakest for mathematics, where differences were only statistically significant for the fourth position on the first measurement occasion and for the third and fourth position in the Grade 10 assessment. In the case of science, significant school type differences were found for positions three and four in the first assessment. In the 10th grade, differences were only significant in the fourth position because the standard errors of TCEs increased due to the smaller number of students working on science items (see Table 1). School type differences in TCEs were largest for reading where differences were significant for positions three and four in the first assessment, and for positions two to four in the second assessment.

Changes in TCEs by School Type. An additional observation made on the basis of the entries in Table 3 is that average TCEs appeared to become stronger in the second assessment in the nonacademic school types. In the case of academic schools, average TCEs remained roughly stable in mathematics, and even appeared to become somewhat weaker in reading. Note, however, that the estimates given in Table 3 are based on TCEs averaged over different clusters in the first and second assessment. Hence, the across-time differences could be partially due to the fact that the clusters less sensitive to PEs might not have been included in the second assessment.

In order to overcome this problem, changes in TCEs were analyzed on the basis of clusters presented on both measurement occasions (Table 4). In the case of mathematics, we considered only the clusters M2, M4, and M7. With the exception of M2, which was not presented in cluster position 2 in Grade 10, M4 and M7 were presented in all positions on both occasions. For science and reading, clusters S1 and R1 were not further considered because they were not administered in the second assessment. As is shown in Table 4, in mathematics, average TCEs remained largely constant in the academic school type, but became somewhat more pronounced in the nonacademic schools, where changes were significant for average TCEs presented in the last position. In science, average TCEs became significantly stronger in the last cluster position in both school types. Finally, for reading, TCEs were consistently more negative on the second occasion in nonacademic school types, but only the effect referring to the third cluster position turned out to be significant. In the academic school track, TCEs were reduced on the second occasion (indicated by positive signs in Table 4), but changes were not significant. Table 4 also displays the differences in the changes of average TCEs between school types. Even though a general picture emerged, indicating that PEs become more accentuated in nonacademic schools (indicated by negative sign of differences), no comparison turned out to be statistically significant at the $p < .05$ level. Only in the case of reading were school-type differences in changes of TCEs rather large and significant at the $p < .10$ level.

Table 3:
Test Context Effects Averaged Across Clusters by Cluster Position, and Differences in Test Context Effects between School Types (Academic – Nonacademic)

	Mathematics			Science			Reading		
	Nonacad.	Academic	Difference	Nonacad.	Academic	Difference	Nonacad.	Academic	Difference
	Est.(*SE*)	*Est.*(*SE*)	*Est.*(*SE*)	*Est.*(*SE*)	*Est.*(*SE*)	*Est.*(*SE*)	*Est.*(*SE*)	*Est.*(*SE*)	*Est.*(*SE*)
Grade 9									
ICP2	-0.05(.03)	0.02(.03)	-0.07(.04)	-0.05(.04)	-0.01(.04)	-0.03(.06)	-0.35(.05)**	-0.21(.06)**	-0.14(.08)
ICP3	-0.19(.03)**	-0.12(.03)**	-0.07(.04)	-0.18(.05)**	-0.03(.04)	-0.15(.06)*	-0.35(.05)**	-0.20(.06)**	-0.15(.08)*
ICP4	-0.27(.03)**	-0.18(.03)**	-0.08(.04)*	-0.35(.04)**	-0.15(.04)**	-0.20(.05)**	-0.75(.04)**	-0.48(.05)**	-0.27(.07)**
Grade 10									
ICP2	-0.11(.05)*	-0.06(.05)	-0.06(.06)	-0.03(.09)	-0.01(.10)	-0.02(.14)	-0.57(.12)**	-0.05(.13)	-0.52(.17)**
ICP3	-0.27(.05)**	-0.11(.05)*	-0.15(.07)*	-0.27(.09)**	-0.10(.10)	-0.17(.14)	-0.72(.12)**	-0.30(.13)*	-0.42(.16)**
ICP4	-0.36(.05)**	-0.18(.05)**	-0.19(.07)**	-0.69(.08)**	-0.34(.08)**	-0.35(.11)**	-0.92(.12)**	-0.33(.13)*	-0.60(.17)**

Notes. ICP = Item cluster position
* $p \leq .05$; ** $p \leq .01$

Table 4:
Changes in Test Context Effects Averaged Across Clusters Presented on Both Occasions by Cluster Position and Differences in Test Context Effects between School Types (Academic – Nonacademic)

	Mathematics			Science			Reading		
	Nonacad.	Academic	Difference	Nonacad.	Academic	Difference	Nonacad.	Academic	Difference
	Est.(*SE*)	*Est*.(*SE*)	*Est*.(*SE*)	*Est*.(*SE*)	*Est*.(*SE*)	*Est*.(*SE*)	*Est*.(*SE*)	*Est*.(*SE*)	*Est*.(*SE*)
ICP2	-0.10(.08)	-0.07(.08)	-0.04(.10)	0.08(.11)	0.00(.11)	0.08(.15)	-0.18(.13)	0.15(.15)	-0.34(.19)
ICP3	0.01(.08)	0.04(.08)	-0.03(.11)	-0.03(.10)	-0.07(.11)	0.04(.15)	-0.32(.13)*	-0.06(.14)	-0.26(.18)
ICP4	-0.16(.07)**	-0.04(.07)	-0.12(.10)	-0.29(.09)**	-0.20(.10)*	-0.09(.13)	-0.13(.14)	0.22(.14)	-0.35(.19)

Notes. ICP = Item cluster position
* $p \leq .05$; ** $p \leq .01$

Taken together, the results reported so far clearly document the existence of TCEs that are apparently in line with PEs. In addition, the different domains assessed in PISA seem to be affected to a different degree by such effects, and effects appeared to vary in the expected direction across school types. Furthermore, the analyses provided some indication that PEs became stronger in the second measurement occasion, but this pattern appeared to be mostly characteristic of the nonacademic school types. However, these results did not appear to be very robust because they had large standard errors.

Consequences of test context effects for assessing group differences

The results reported in the previous section indicate that the TCEs investigated could have consequences for assessing group differences. The finding that the TCEs were, on average, larger in nonacademic school types implies that ignoring these effects is likely to result in larger achievement differences in favor of academic schools on each occasion. In addition, the finding that TCEs were more pronounced on the second occasion in the nonacademic schools could mean that ignoring TCEs might result in group differences in achievement gains that favor the academic group more strongly.

Cross-Sectional School Type Differences in Proficiency Levels. Table 5 reports the estimates of cross-sectional latent means, standard deviations, and group differences, as well as the estimates of gains in proficiency taken from different IRT models with and without TCEs. Group differences in means in favor of academic school types were generally large in all domains. Accounting for TCEs generally reduced cross-sectional mean differences, with reductions being somewhat smaller on the first measurement occasion. In mathematics, the relative size of mean differences on the first occasion was reduced by 3% after including TCEs, and was reduced by 6% on the second occasion of measurement. In science, group differences were reduced by 7% and 10% on the first and the second occasion, respectively. The largest differences were found for reading, with a reduction of 10% on the first, and 28% on the second occasion.

Longitudinal School Type Differences in Proficiency Levels. Although many of these changes might appear to have rather trivial consequences for assessing group differences in achievement gains, the entries in Table 5 show that only the results for mathematics were generally robust against TCEs. The IRT models that did not account for TCEs estimated statistically significant negative achievement gains for the nonacademic schools in science and reading, whereas the corresponding estimates were essentially 0 in the IRT models accounting for TCEs. However, the inclusion of TCEs had negligible consequences for the growth estimate in academic schools. This pattern reflects the fact that the size of TCEs did not change much in this group across occasions (Tables 3 and 4).

Accounting for TCEs affected the results of group comparisons in changes of average proficiency levels. The difference between the two models envisaged was rather small in the case of mathematics, but the IRT model with TCEs resulted in the group difference in growth being 28% smaller than the result achieved by the model ignoring TCEs. In

Table 5:
Latent Means, Standard Deviations, and Mean Differences Derived by IRT Models Excluding (2-Dim-1PL) and Including (2-Dim-1PL+Booklet) Booklet Effects

	2-Dim-1PL					2-Dim-1PL+Booklet				
	Nonacademic		Academic		Difference	Nonacademic		Academic		Difference
	M(*SE*)	*SD*	*M*(*SE*)	*SD*	*D*(*SE*)	*M*(*SE*)	*SD*	*M*(*SE*)	*SD*	*D*(*SE*)
Mathematics										
Grade 9	0.00(---)	0.82	1.27(.03)**	0.83	1.27(.03)**	0.00(---)	0.82	1.23(.04)**	0.83	1.23(.04)**
Grade 10	0.01(.02)	0.95	1.46(.03)**	0.97	1.45(.04)**	0.06(.04)	0.94	1.42(.05)**	0.96	1.36(.05)**
Difference	0.01(.02)	0.51	0.19(.03)**	0.44	0.18(.03)**	0.06(.04)	0.49	0.19(.05)**	0.44	0.13(.06)*
Science										
Grade 9	0.00(---)	0.79	0.95(.05)**	0.66	0.95(.05)**	0.00(---)	0.79	0.88(.06)**	0.66	0.88(.06)**
Grade 10	-0.12(.04)**	0.97	1.20(.06)**	0.90	1.31(.07)**	-0.03(.07)	0.95	1.16(.09)**	0.88	1.18(.10)**
Difference	-0.12(.04)**	0.59	0.24(.04)**	0.54	0.36(.05)**	-0.03(.07)	0.54	0.28(.07)**	0.52	0.31(.10)**
Reading										
Grade 9	0.00(---)	0.84	1.10(.03)**	0.74	1.10(.03)**	0.00(---)	0.83	0.99(.06)**	0.73	0.99(.06)**
Grade 10	-0.23(.04)**	1.15	1.16(.04)**	0.91	1.39(.06)**	0.02(.10)	1.11	1.02(.11)**	0.90	1.00(.13)**
Difference	-0.23(.04)**	0.88	0.06(.04)	0.70	0.29(.06)**	0.02(.10)	0.79	0.03(.11)	0.69	0.01(.13)

science, group differences were reduced by 14% after accounting for TCEs, whereas in reading, the differences almost completely vanished (97%) once TCEs were controlled for. The pattern of results regarding group differences in average change is mainly due to the pronounced effects of TCEs on change estimates in nonacademic schools.

Summary and discussion

In the present article, we investigated TCEs in the longitudinal extension of the PISA 2012 assessment in Germany. For this purpose, an IRT model including TCEs operating on the level of item clusters was proposed. In this approach, TCEs are defined as the difference between how difficult items are when they are presented in a given cluster position relative to when they are presented in the first position. The suggested approach is flexible as it allows TCEs to vary across clusters without assuming a predefined functional form of effects. Hence, our approach is sensitive to many kinds of TCEs, including the effects of positions and the ordering of domains in different test booklets, among others.

The application of the model to the longitudinal PISA 2012-2013 assessment revealed that the pattern of TCEs largely reflected PEs, as indicated by the nearly monotone decline in the probability of correct responses, the later the clusters were presented in a test booklet. The fact that the achievement tests used in PISA are prone to PEs is not new (e.g., Debeer & Janssen, 2103; Debeer et al., 2014; Hartig & Buchholz, 2012; Wu, 2010). However, our findings extend previous research in many respects. First, our results point out that the largest changes in achievement decrements during testing took place in the second half (reading) or in the last quarter (mathematics and science) of the test, indicating that PEs apparently do not operate in a linear fashion, as assumed in most IRT approaches used for estimating PEs (e.g., Debeer & Janssen, 2013; Hohensinn et al., 2008). Second, our analyses provide strong evidence that students from different school types are affected by PEs to a different degree, with effects being most strongly pronounced in nonacademic schools. This finding is in line with other results showing that PEs are stronger in lower achieving groups (Debeer et al., 2014; Hartig & Buchholz, 2012). Third, our results provide some indication that PEs tended to become more accentuated in the retest assessment, while this pattern was most pronounced for the reading domain in nonacademic schools. This result is in line with our speculations that retesting students with a long test (about two hours) in a low-stakes condition might be experienced more adversely by lower achieving students. Fourth, the findings clearly document that TCEs affect the results of group comparisons. Cross-sectional differences in the first assessment were only mildly affected because mean differences between school types were very large. However, results for the second occasion of measurement were more strongly impacted by PEs because group differences in PEs tended to become stronger. In the present case, the impact of PEs was strong enough to result in negative estimates of growth in nonacademic schools in two out of three domains tested (reading and science) when TCEs were ignored. Such a result is not only counterintuitive, but stands in sharp contrast to findings documenting small or absent achievement gains from the 9th to the 10th grade (e.g., Bloom, Hill, Black, & Lipsey, 2008).

Implications for large-scale assessment of student achievement

Our findings strongly suggest that TCEs in general and PEs in particular should be taken seriously in large-scale studies of student achievement. However, assessing and controlling such effects requires optimal designs and suitable statistical models. Regarding the test design, the sound identification of PEs requires items to be presented in varying positions. Weirich, Hecht, and Böhme (2014) have investigated this issue and shown that test designs should optimally balance the presentation of each test part (e.g., clusters) across positions. However, by introducing the notion that other types of TCEs might operate alongside PEs, test designs should either allow each sort of effects to be identified, or they should focus on estimating TCEs that represent a compound of a multitude of different effects.

In our view, the assessment of pure PEs is not a realistic option for most large-scale studies. Such designs would either require some form of focused testing that is restricted to one domain, or booklets in which positions are fully crossed with domain orders. The first design option counteracts many intended purposes of large-scale studies, such as assessing the correlation between domains. The second design option is not realistic because it requires a very large number of booklets, which is impractical for most applied settings.

Hence, an approach oriented towards identifying compounds of different contextual effects (i.e., TCEs) appears the most reasonable strategy in large-scale studies of student achievement. We believe that the suggested model is capable of providing solid results, although the quality of the results clearly depends on the sample size. In contrast to the design options of focused testing, and full crossing of positions and domains orders, test designs suitable for large-scale assessments might be quite easily optimized to increase the stability of results. Most importantly, care should be taken that the number of students providing responses for the first cluster position is large in each domain. This is because the first position is used as an anchoring point for assessing the students' proficiencies. The smaller the number of students responding to the first position, the less precise the estimation of the latent means and their changes will be. This issue became evident in the present application, as the standard errors of mean parameters and their changes were relatively small for the mathematics domain, which was assessed with the largest number of clusters, and were the largest for reading, which was most weakly anchored on the second measurement occasion.

Note, however, that the relatively large standard errors in reading do not mean that the estimates provided by our model are biased. Biases should rather result as consequences of a misshapen random allocation of students to booklets and/or as a violation of the measurement invariance assumption imposed on the item parameters. However, both assumptions can be checked empirically, and the model can be modified if they are violated. Such modifications could involve including covariates in order to account for group differences due to a suboptimal random allocation, and relaxing the invariance constraints on item parameters because such deviations could also be seen as an indication of TCEs being at work on the item level. In the present article, we carefully checked this possibility and found no evidence for such deviations.

Limitations and future directions

We developed an IRT model to assess TCEs and study their effects in estimating cross-sectional and longitudinal group differences in average proficiencies. Although the ideas on which our statistical approach is based are simple, we believe that the presented approach provides further insights into how TCEs operate in longitudinal large scale assessments. Nevertheless, as we have outlined before, the assessment designs suitable for the proposed IRT model can be optimized in order to provide more reliable results (cf. Weirich et al., 2014). Indeed, in terms of the booklet designs employed, the model was found to provide estimates of low reliability (i.e., large standard errors) in the case of reading. This finding is not a failure of the model itself, but rather a consequence of the sparse data available for this domain. However, more research is clearly needed to evaluate the accuracy of the standard errors found by our model under various conditions.

In addition, the model presented can be extended in various ways to make it better suited to other research questions. For example, applications might call for the inclusion of continuous predictors of TCEs; this request could be accommodated by including interactions between booklet indicators and continuous covariates. For example, researchers might be interested in whether cognitive capacities or measures of test motivation are related to TCEs. Other applications might call for more complex measurement models, such as the 2PL, which can be easily accommodated by the model as well. Further extensions could be devoted to the examination of individual differences in TCEs. However, in contrast to IRT models for randomly varying PEs (e.g., Debeer & Janssen, 2013), the inclusion of continuously distributed TCEs appears to be a challenge when TCEs are assumed to be specific for each combination of booklets and clusters. An alternative route might be to combine the model with a latent class model. This would make it possible to identify groups of students that are differently affected by TCEs.

Finally, it should be noted that, in the present application, we inspected how TCEs affected the probabilities of correct responses. However, TCEs could theoretically also be related to other outcomes, such as the probability of missing item responses. For example, PEs might be linked to the probability of not reaching the end of the test (e.g., Glas & Pimentel, 2008), and DOEs could be candidates of increasing probabilities of omissions (e.g., Holman & Glas, 2005). The model presented in this article can be extended in order to study whether TCEs affect the probability of omitting and/or not reaching items in the test. However, more work is clearly needed to accommodate the proposed model for such situations, because the literature suggests that missing data is better modeled on the level of items than on the level of item clusters, as undertaken in our model.

References

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*, 289-328.

Brennan, R. L. (1992). The Context of Context Effects. *Applied Measurement in Education, 5*, 225-264.

Debeer, D., & Janssen, R. (2013). Modeling item position effects within an IRT framework. *Journal of Educational Measurement, 50*, 164-185.

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.

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*, 23-45.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica*, *37*, 359-374.

Frey, A., Hartig, J. & Rupp, A. (2009). Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice, 28*, 39-53.

Glas, C. A., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological* Measurement, 68, 907-922.

Harris, D. (1991). Effects of passage and item scrambling on equating relationships. *Applied Psychological Measurement, 15*, 247-256.

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.

Hauser, R. M., & Goldberger, A. S. (1971). The treatment of unobservable variables in path analysis. *Sociological Methodology*, *3*, 81-117.

Heine, J.-H., Nagy, G., Meinck, S., Zühlke, O., & Mang, J. (2016). Empirische Grundlagen und Stichprobenausfall im PISA-Längsschnitt 2012-2013 [Empirical background and sample drop-out in the longitudinal PISA 2012-2013 study]. *Manuscript submitted for publication*.

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science*, *50*(3), 391.

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*, 1-17.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*, 387-413.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Liu, J., & Dorans, N. J. (2012). Assessing the practical equivalence of conversions when measurement conditions change. *Journal of Educational Measurement*, *49*, 101-115.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*, 38–60.

Moses, T., Yang, W. L., & Wilson, C. (2007). Using kernel equating to assess item order effects on test scores. *Journal of Educational Measurement*, *44*, 157-178.

Muthén, L.K., & Muthén, B.O. (1998-2012). *Mplus user's guide. Seventh edition*. Los Angeles, CA: Muthén & Muthén.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., Estabrook, R., Bates, T. C., Maes, H. H., & Boker, S. M. (2015). OpneMX: Extended structural equation and statistical modeling. *Psychometrika, 81*, 535-549.

Organization for Economic Cooperation and Development (OECD) (2014). *PISA 2012 technical report*. Paris: OECD Publishing.

Qian, J. (2014). An Investigation of Position Effects in Large-Scale Writing Assessments. *Applied Psychological Measurement*, Published online before print June 3, 2014.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, H.-G. R., Rost, J., Schiefele, U. (Hrsg.). (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente*. Münster: Waxmann Verlag.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Kopenhagen: Paedagogiske Institut.

Rindskopf, D. (1984). Using phantom and imaginary latent variables to parameterize constraints in linear structural models. *Psychometrika*, *49*, 37-47.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76*, 318-336.

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement*, *38*(7), 535-548.

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. Educational Measurement: *Issues and Practice, 29*, 15–27.

Wu, E. J. C., & Bentler, P. M. (2011). EQSIRT: A user-friendly IRT program (Computer software). *Encino, CA: Multivariate Software*.

Appendix

Mplus Syntax for the 2-Dim-1PL+Booklet Model Applied to the Science Data

```
Title:       2-DIM-1PL+Booklet Model

Data:        file is pisa_science.dat;

Variable:    names are

             sid   ! School-ID

             0 = academic 1 = nonacademic
```

```
            stype ! School-type
            ab07 ab08 ab01 ab10 ab03 ab05 ab12 ab13
            ! Booklet indicators T1
            bb01 bb03 bb14 bb17 bb11 bb18
            ! Booklet indicators T2
            AS1_01-AS1_18  ! Items in cluster 1 at T1
            AS2_01-AS2_18  ! Items in cluster 2 at T1
            AS3_01-AS3_16  ! Items in cluster 3 at T1
            BS2_01-BS2_18  ! Items in cluster 2 at T2
            BS3_01-BS3_16; ! Items in cluster 3 at T2

            categorical are
            AS1_11 AS2_18 AS3_11 BS2_18 BS3_11 (pcm)
            !Partial-credit items
            AS1_01-AS1_10 AS1_12-AS2_17 AS3_01-AS3_10 AS3_12-BS2_17
            BS3_01-BS3_10 BS3_12-BS3_16;
            missing are all (7,8,9);

            ! Known latent class variable for multigroup analysis
            classes = c(2);
            knownclass = c (stype = 0 stype = 1);

            cluster is sid; ! specification of clustering variable

Analysis:   ! Accounting for clustering of observations
            type = complex mixture;
            algorithm = integration;
            link = logit;

Model:      ! Specification of the general model structure
            %overall%

            ! Definition of proficiency variables with unit loadings
            sci1 by AS1_01-AS3_16@1;
            sci2 by BS2_01-BS3_16@1;
```

```
! Definition of node-variables capturing TCEs
ac1 by AS1_01-AS1_18@1;
ac2 by AS2_01-AS2_18@1;
ac3 by AS3_01-AS3_16@1;
bc2 by BS2_01-BS2_18@1;
bc3 by BS3_01-BS3_16@1;

! Constraints of node-variables to have zero means,
! zero variances, and zero covariances
[ac1-bc3@0];
ac1-bc3@0;
ac1-bc3 with ac1-bc3@0 sci1@0 sci2@0;

! Regression of node-variables on booklet indicators
ac1 on ab08 ab03 ab05;
ac2 on ab07 ab01 ab13;
ac3 on ab01 ab10 ab12;
bc2 on bb01 bb11 bb18;
bc3 on bb03 bb14 bb17;

! Equality constraints on item difficulties

! Dichotomous items of item cluster S2 at T1
[as2_01$1-as2_17$1] (i2_01-i2_17);
! Partial credit items of item cluster S2 at T1
[as2_18$1] (i2_181);
[as2_18$2] (i2_182);

! Dichotomous items of item cluster S3 at T1
[as3_01$1-as3_10$1] (i3_01-i3_10);
[as3_12$1-as3_16$1] (i3_12-i3_16);
! Partial credit items of item cluster S3 at T1
[as3_11$1] (i3_111);
[as3_11$2] (i3_112);

! Dichotomous items of item cluster S2 at T2
[bs2_01$1-bs2_17$1] (i2_01-i2_17);
```

```
             ! Partial credit items of item cluster S2 at T2
             [bs2_18$1] (i2_181);
             [bs2_18$2] (i2_182);

             ! Dichotomous items of item cluster S3 at T2
             [bs3_01$1-bs3_10$1] (i3_01-i3_10);
             [bs3_12$1-bs3_16$1] (i3_12-i3_16);
             ! Partial credit items of item cluster S3 at T2
             [bs3_11$1] (i3_111);
             [bs3_11$2] (i3_112);

             ! Model part for academic schools (class 1)
             %c#1%
             ! Means of proficiency variables
             [sci1 sci2];

             ! Covariance structure of proficiency variables
             sci1 WITH sci2;
             sci1 sci2;

             ! Fixed mean Structure of node variables
             [ac1-bc3@0];

             ! Regression of node-variables on booklet indicators
             ac1 on ab08 ab03 ab05;
             ac2 on ab07 ab01 ab13;
             ac3 on ab01 ab10 ab12;
             bc2 on bb01 bb11 bb18;
             bc3 on bb03 bb14 bb17;

             ! Model part for nonacademic schools (class 2)
             %c#2%
             ! Overriding the default constraint that all
             ! latent variable means are 0 in the reference class
             [sci1@0 sci2];

Output:      tech1 tech8;
```