Modeling Booklet Effects for Nonequivalent Group Designs in Large-Scale Assessment

Educational and Psychological
Measurement

2015, Vol. 75(4) 568–584

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414554219

epm.sagepub.com

Martin Hecht¹, Sebastian Weirich¹,
Thilo Siegle¹, and Andreas Frey²

Abstract

Multiple matrix designs are commonly used in large-scale assessments to distribute test items to students. These designs comprise several booklets, each containing a subset of the complete item pool. Besides reducing the test burden of individual students, using various booklets allows aligning the difficulty of the presented items to the assumed performance level of examined subgroups. While this may improve measurement precision and students' test-taking motivation, using several booklets might influence response behavior and thus constitute a potential source of unwanted variation. To provide guidance to identify and model booklet effects, this study presents statistical models accounting for booklet effects and applies these models in a large-scale assessment setting. Three models are derived from the Rasch model employing the generalized linear mixed models framework. The models were applied to data from a national educational standards assessment study for scientific competence. A total of 1,021 items were compiled to 74 booklets distributed to a sample of 9,044 students of Grades 9 and 10. The results revealed a small but nonnegligible booklet effect. For further large-scale assessment studies, it is recommended to examine whether booklet effects occur and to adequately account for them in the subsequent analyses where necessary.

Keywords

testing, large-scale assessment, generalized linear mixed models (GLMM), context effects, multiple matrix sampling, nonequivalent groups

---

¹Humboldt-Universität zu Berlin, Berlin, Germany

²Friedrich Schiller University Jena, Jena, Germany

Corresponding Author:

Martin Hecht, Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.

Email: martin.hecht@iqb.hu-berlin.de

Large-scale assessments of student achievement generally strive to measure and compare the competence of students in one or several domains of interest. Usually large collections of items are needed to comprehensively cover all relevant aspects of the competences at stake. In the Programme for International Student Assessment (PISA), for example, about 150 to 200 items are employed to measure students' literacy in reading, mathematics, and science in each assessment. To keep the workload within acceptable boundaries, only subsets of the available items are presented to the individual students. These item subsets are usually called *booklets* in the case of paper-and-pencil tests, or more generally *test forms*. Booklets are assembled using rules and techniques such as multiple matrix sampling (Frey, Hartig, & Rupp, 2009; Gonzalez & Rutkowski, 2010; Rutkowski, Gonzales, von Davier, & Zhou, 2014; Shoemaker, 1971). One consequence of this assembly process is that different booklets contain different sets of items and can thus vary in their properties like their average item difficulty or number of items.

Based on the item difficulty known *a priori* to the booklet assembly process, the assumed booklet difficulty can be varied systematically across booklets, which will be referred to as *a priori booklet difficulty* in the following text. This variation can be advantageous because it allows for aligning the *a priori* booklet difficulty to the competence level of examined groups of students. This is advantageous because in item response theory (IRT) models, measurement precision is not constant across the measurement scale but varies depending on the availability of items that measure in a particular region of the scale (e.g., Embretson & Reise, 2000). Thus, assembling booklets with a different average item difficulty and distributing them to competence-matched subsamples increases measurement precision since every subpopulation receives items with higher statistical information compared with a random selection of items.

Second, especially in low-stakes testing, competence measures are prone to be impaired by motivational and emotional (e.g., effort, boredom, anxiety) influences (Cole, Bergin, & Whittaker, 2008; Eklöf, 2010; Wise & DeMars, 2005; Wolf & Smith, 2005). For instance, for students with low test-taking motivation, the observed competence is lower than his or her actual competence (Wise & DeMars, 2005). Moreover, the individual effort spent working on test items depends on the difference between a student's competence and the difficulty of the booklet used (Asseburg & Frey, 2013). Furthermore, students who are confronted with far too easy or far too hard booklets might become bored or frustrated, and therefore they might perform differently than they would have if booklets with a more appropriate difficulty level had been administered. van der Linden, Veldkamp, and Carlson (2004) summarize this problem:

... careless, unmotivated test taking is a known problem in such studies. Careless behavior is expected to increase if the items are much too easy, and motivation decreases if they are much too difficult for the students. It may therefore pay off to design the assessment

booklets such that each population receives items with a probability of success as close as possible to an optimum.... (p. 321)

Indeed, the optimization of low-stakes tests to avoid negative motivational effects is a central idea behind computerized adaptive testing (Linacre, 2000; van der Linden & Glas, 2010; Wainer, 2000). It was frequently argued that adapting the test to the individual competence level fosters test-taking motivation, and therefore the construct-irrelevant variance in the competence measurement is reduced. Critically examining this claim, Wise (2014) concludes that “there is evidence that the targeting of item difficulty to examinee proficiency brings with it modest motivational benefits that can improve individual score validity” (p. 13).

Alignment of *a priori* booklet difficulty and student competence level is applied in several large-scale assessments. In PISA (OECD, 2005, 2009, 2012) booklets that contain less and easier items can be given to students with special educational needs. The same approach is applied in national assessments in Germany (e.g., Hecht, Roppelt, & Siegle, 2013; Weirich, Haag, & Roppelt, 2012).

Nevertheless, varying *a priori* booklet difficulty need not always be advantageous; it also can have drawbacks. One potential disadvantage is based on the assumption of the statistical models that are commonly applied in large-scale assessments—for instance, the Rasch model or the two-parameter logistic model (2PL)—that features of the measurement instrument do not have an effect on the responses given to the test items. The Rasch model, for example, simply states that the probability $P(Y_{jib} = 1)$ of a student j to correctly solve an item i in booklet b is dependent on the competence θ_j of the student and an item difficulty parameter β_i . Note that the probability to solve the item j is typically not considered to depend on the empirical booklet difficulty, denoted as γ_b in the remainder of this article. The Rasch model estimates might be biased if the empirical booklet difficulty has an additional effect, beyond items and students, on the probability of giving a correct response. That such context effects can occur with deleterious consequences is a well-documented phenomenon. Leary and Dorans (1985) provide an extensive literature review and conclude, “In sum, the research literature has found evidence of context effects” (p. 410). A popular example of critical consequences of context effects is the NAEP 1985-1986 reading anomaly (Beaton, 1988). Briefly, from one wave to another, students’ reading competence decreased in an unexpected magnitude. This finding was so puzzling that reporting of results was postponed. Further investigations showed that “the changes in assessment booklets and procedures that were introduced in 1986 had a substantial and unpredictable effect on the estimates of performance” (Beaton & Zwick, 1990, p. xi). Thus, multiple matrix sampling that produces booklets with differing properties should be used with caution. In fact, Lord (1962) had already warned, “Any practical application of the item-sampling method thus involves the further assumption that the

examinees' performance on the items is not too greatly affected by the context in which they are administered'' (p. 263).

But how can a potential bias of the parameter estimates of IRT models due to potential booklet effects be avoided? One strategy is to keep all potentially influencing booklet properties constant. This results in using the same booklet for all students, which is not feasible in large-scale assessments and thwarts the central idea behind multiple matrix sampling. Another strategy would be to balance potential booklet effects by distributing all booklets randomly to all students. In this case, potential effects of booklets on parameter estimates of interest are held constant between groups as long as the groups are not too small. Because the statistics of interest in typical large-scale assessments are based on large groups, booklets with varying properties can be regarded as unproblematic in that area. However, if the *a priori* booklet difficulty should be tailored to the competence level of a subgroup to enhance measurement precision, balancing only works within subgroups but not overall. A solution to this problem is to apply statistical models that explicitly include booklet effects to examine whether they occur in such a substantive manner that commonly used IRT models might be inappropriate to analyze the data at hand.

One possibility to include booklet effects in the statistical model is by employing the generalized linear mixed models (GLMM; De Boeck, 2008; Wilson & De Boeck, 2004) framework. Specifically, a booklet parameter γ_b may be added to investigate whether the probability of correctly responding to an item is influenced by the booklet used. The interpretation is rather simple: while θ_j and β_i describe in what way the variability of students and items is responsible for the variability of responses, γ_b describes in what way the variability of booklets is *additionally* responsible for the variability of responses.

However, the modeling of unit estimates or their distribution characteristics (e.g., the variance of a potential booklet effect) is usually just the first step. Of major interest is not only to estimate the booklet variance but to explain this variance by further covariates to answer the "why question". While item properties and person properties have been previously investigated (e.g., Freedle & Kostin, 1993; Leucht, Harsch, Pant, & Köller, 2012), research on booklet properties is sparse. Booklet properties that might influence the probability to solve an item are either located at the surface (e.g., text font, size or color) or produced by the compilation of booklet-specific item sets. When items are assigned to booklets, booklet properties can be derived from the aggregation of item properties. For instance, item format (i.e., multiple-choice vs. constructed-response) can be aggregated to the percentage of multiple-choice items in a booklet, and thus this aggregated variable can be regarded as a booklet property. This booklet property might then explain the variance of the empirical booklet difficulty with booklets containing relatively more multiple-choice items probably being easier than booklets with relatively more constructed-response items. Another example is the number of items per booklet. Since items usually vary in their allocated completion time, sometimes booklets with varying item numbers are used in one single study to reach a constant testing time per booklet. Nevertheless, working on a

larger number of items might affect response behavior compared with answering a smaller number of items—more items imply dealing with and thinking about more issues, making more decisions, or writing more responses. Thus, booklets with more items might empirically turn out to be more difficult.

Research Scope

Two research goals are pursued: The first is to introduce models that can be used to model booklet effects. These models will be formulated within the GLMM framework. Four models with increasing complexity are examined. The first is the Rasch model. The second is suitable if the booklets are distributed randomly to the students. The third is appropriate if the distribution of booklets depends on a student covariate (e.g., school track, or some other competence-related grouping variable). The fourth includes booklet properties to examine whether these are related to estimated booklet effects. The second goal is to illustrate that the proposed models can productively be used to examine booklet effects in a large-scale assessment setting.

Models

The three booklet models introduced in this study are derived from the Rasch model. Its mathematical formulation for $j = 1, \dots, J$ persons and $i = 1, \dots, I$ items is

$$P(Y_{ji} = 1) = \text{logit}^{-1} (\alpha_0 + \theta_j + \beta_i) \quad (1)$$

with the assumptions of Y_{ji} being Bernoulli-distributed with location $P(Y_{ji} = 1)$, and θ_j and β_i being normally distributed with zero means and variances σ_θ^2 and σ_β^2 , respectively. The central postulation of the Rasch model is that the dichotomous response (i.e., either correctly or incorrectly solving an item) solely depends on the competence of a person, θ_j , and the easiness of the item, β_i . To account for the dichotomous nature of the outcomes, the logit link is applied to map the mean of the Bernoulli-distributed responses onto the latent continuous scale (e.g., De Boeck & Wilson, 2004). We use the random person–random item (RPRI; De Boeck, 2008) formulation of the Rasch model, assuming that persons and items are sampled from a normally distributed population of persons and items, respectively. When assuming θ_j and β_i as random effects, an intercept α_0 needs to be added to model the mean difference between persons and items. Furthermore, in the original formulation of the Rasch model item parameters β_i are subtracted. If item parameters are added, the interpretation changes from item *difficulty* to item *easiness*. To remain consistent with the parameterization in the *lme4* package that is used for estimation we will adopt the “+” parameterization and corresponding interpretations throughout the rest of the article. In the Rasch model, the parameter estimates are usually called logits since they are logarithmized quotients of the probability of a person to correctly respond and incorrectly respond to an item.

The first model (booklet model) is an extension of the Rasch model in Equation 1, supplemented by a normally distributed booklet parameter γ_b for each booklet $b = 1, \dots, B$:

$$P(Y_{jib} = 1) = \text{logit}^{-1} (\alpha_0 + \theta_j + \beta_i + \gamma_b) \quad (2)$$

The booklet parameters are positively parameterized and thus can be interpreted as booklet *easiness*. Analogous to person and item parameters we assume that there is a universe of booklets that a sample is taken from, and that the distribution of booklet easiness is normal with zero mean and variance σ_γ^2 . Thus, booklets are modeled as a random effect. The booklet model should produce unbiased person estimates if booklets are distributed to students randomly. However, if it is intended to assign specific booklets to several groups of students that differ in competence (i.e., nonrandom assignment), this model's estimated parameters will most certainly deviate from their true parameters due to the failure to account for sources of variation. For the sake of simplicity, we will subsume such effects in the term *bias*. Under nonrandom assignment, this potential bias is mainly due to the fact that booklets are not linked with each other, since every student only obtains one booklet and not several. Thus, each booklet parameter is estimated in reference to the specific subgroup of students that received that booklet but not in reference to the entire sample. Booklets that are distributed to more competent students then will appear easier, and booklets that are distributed to less able students will appear more difficult. A way to control for subgroup differences in the estimated booklet effects is to include the nonequivalent groups into the booklet model. Hereby, booklets and the nonequivalent groups need to be (at least partially) crossed, that is, at least some booklets must be delivered to more than one group. The previous booklet model is now extended by adding a group effect δ_s and a booklet $\times$ group interaction ε_{bs} , which results in the nonequivalent groups booklet (NEGB) model:

$$P(Y_{jibs} = 1) = \text{logit}^{-1} (\alpha_0 + \theta_j + \beta_i + \gamma_b + \delta_s + \varepsilon_{bs}) \quad (3)$$

In a next step, booklet properties can be added to the model as fixed effects to predict the booklet variance by adapting the approach for modelling property covariates described in de Boeck et al. (2011). For this purpose γ_b is decomposed into the sum of values of a booklet property matrix $\mathbf{X}$ that contains values of booklet property k ($k = 1, \dots, K$) for each case and a fixed effect κ_k for each property k : $\gamma_b = \sum_{k=1}^K \kappa_k X_k + \nu_b$. To account for a possibly imperfect prediction, a residual term ν_b with an assumed normal distribution with zero mean and variance σ_ν^2 is added. This yields the nonequivalent groups booklet properties (NEGBP) model:

$$P(Y_{jibs} = 1) = \text{logit}^{-1} \left(\alpha_0 + \theta_j + \beta_i + \sum_{k=1}^K \kappa_k X_k + \nu_b + \delta_s + \varepsilon_{bs} \right) \quad (4)$$

A small example to illustrate the data structure for this kind of models is given in Table 1 as data in long format and in Figure 1 as schematic visualization. In our

Table 1. Example of Data Structure in Long Format.

Student	Group	Booklet	Item	N Item	Y_{jibs}
1	1	1	1	3	1
2	1	1	2	3	1
3	1	1	3	3	0
4	1	2	1	6	1
5	1	2	2	6	0
6	1	2	3	6	0
7	1	2	4	6	0
8	1	2	5	6	1
9	1	2	6	6	0
10	2	2	1	6	1
11	2	2	2	6	1
12	2	2	3	6	1
13	2	2	4	6	1
14	2	2	5	6	0
15	2	2	6	6	1
16	2	3	4	3	0
17	2	3	5	3	1
18	2	3	6	3	1

Note. For each student only one of several rows is shown to improve clarity. “N Item” denotes the number of items in a particular booklet. Y_{jibs} is the dichotomous response of student j in group s on item i in booklet b indicating a correct (= 1) or incorrect (= 0) response.

Figure 1. Visualization of the study design and the implied data structure.

Note. Students belong to one of two groups (e.g., school tracks). A particular booklet is distributed to either Group 1, Group 2, or both. Items and booklets are partially crossed, that is, each item is contained in several booklets, but not in all.

example there are 18 students. Each student belongs to one of two school tracks. If these groups differ in mean students’ competence they are called nonequivalent. Concerning the measurement instruments, there are six items that are assigned to

three booklets. The test design is incomplete since each item is not included in each booklet. For instance, item one is part of booklet one and two, but not of booklet three. In other words, items and booklets are only partially, and thus not completely, crossed. Each student receives one booklet. The assignment of booklets is not completely but conditionally at random, that is, it differs systematically between the nonequivalent groups (e.g., school track). Students from Group 1 (e.g., intermediate track) are randomly assigned to one booklet from the subset of booklets one and two while students from Group 2 (e.g., academic track) will randomly receive either booklet two or three. As an exemplary booklet property, the number of items per booklet is given in the last row. Each booklet contains a distinct number of items that might not be constant across booklets, and thus this booklet property can be related to the booklet effect. The last row of the exemplary data set, Y_{jibs} , is the dichotomous response of student j in group s on item i in booklet b indicating a correct (= 1) or incorrect (= 0) response. This is the dependent variable in all models.

Method

In this section the test design and the student sample from a large-scale assessment study is described. Although this study was not designed to investigate booklet effects, it is typical regarding objectives and methods of large-scale assessment studies and allows the modeling of booklet effects. Although the test design is a commonly used multiple matrix sampling design, some special requirements were also necessary. For this reason, the design assembly procedures might appear somewhat complicated and less relevant. Indeed, a much simpler design would have been sufficient to simply model booklet effects.

Study Design

The data used to illustrate the modeling of booklet effects originates from a large-scale assessment in Germany. The test contained a total of 1,021 items to measure science competence as defined in the German Science Education Standards (Kremer et al., 2012; for details on the development of these standards, see Neumann, Fischer, & Kauertz, 2010). For every item, an empirical estimate of easiness was available from a pilot study. Therefore, it was possible to construct booklets with varying a priori easiness. Since these item parameters are normalized on the nonrepresentative student sample from the pilot study with an average competence lying above the mean of the population, the mean of the item easiness parameters is not zero but $M = 0.28$ ($SD = 1.28$). This scaling is of less importance for the current research because the relevant feature was the generation of variance concerning booklet easiness. In a first step, items were grouped into a total number of 109 disjoint blocks differing in average easiness. Hence, items were nested in blocks. The time allotted for each block was 20 minutes. A priori block easiness was calculated as the mean of item easiness from the pilot study. To keep matters manageable, the

blocks were grouped to 35 easy blocks ($M_B = 1.00$ logits), 38 medium blocks ($M_B = 0.60$), and 36 hard blocks ($M_B = 0.06$). Each of these 109 blocks was used four times in the design; thus, 436 block instances had to be assigned to booklets. Since the testing time of each student was set to 2 hours, each booklet contained six blocks.

Analogous to blocks, three booklet categories were defined: easy, medium, and hard. The hard booklets contained only hard blocks, the easy booklets contained only easy and medium blocks, while the medium booklets were assembled of blocks of all easiness categories. This assembling scheme yielded 74 booklets in total: 25 easy booklets (min = 0.59 logits, max = 1.05, $M = 0.78$), 37 medium booklets (min = 0.36, max = 0.72, $M = 0.55$), and 12 hard booklets (min = 0.01, max = 0.18, $M = 0.06$). For descriptive purposes, a priori booklet easiness values were obtained by calculating the mean of the item easiness resulting in a mean of $M = 0.55$ logits ($SD = 0.26$) and a range of 1.04 logits.

The target population consisted of 9th and 10th grade students. In Germany, the school system varies substantially between the federal states. Still, roughly two school tracks can be differentiated: academic track schools (whose students aspire an academic career) and intermediate track schools (whose students mostly pursue a vocational career). The selection to these two school types generally depends on the achieved grades of the students in primary school. Thus, on average, the mean competence level of the students differs between these school tracks. This competence difference was taken into account in our study design by assigning easier booklets to intermediate track students and more difficult booklets to academic track students. However, *within* both school tracks, booklets were randomly assigned to students. Intermediate track students received either an easy or a medium booklet at random while academic track students received either a medium or a hard booklet at random. This boils down to 62 of the overall 74 booklets with an easiness mean of $M = 0.64$ logits ($SD = 0.16$) being distributed to intermediate track students and 49 booklets ($M = 0.43$, $SD = 0.23$) being distributed to academic track students. In other words, all items have been deployed in both tracks, but intermediate track students received booklets that were 0.21 logits easier on average.

Data Collection

A total of 9,044 9th graders (61.52%) and 10th graders (38.48%) with an average age of $M = 15.66$ years ($SD = 0.81$) were sampled representatively for Germany from intermediate (58.77%) and academic track schools (41.23%). A total of 51.3% of the students were female. In most cases, participation was mandatory; students were neither rewarded nor graded. Data were collected in the spring of 2011. The total testing time was 3.5 hours including breaks. The testing time for the performance items was 2 hours.

Data Analysis

Four generalized linear mixed models were consecutively estimated using the R package *lme4* (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2014): the

Rasch model (Equation 1), the booklet model (Equation 2), the NEGB model (Equation 3), and the NEGBP model (Equation 4). The distributions of the parameters included in the four models were assumed as described above. The nonequivalent groups variable in our design is school track with $S = 2$ groups (i.e., academic track and intermediate track). We chose to use effect coding (i.e., code 1 for academic track and code -1 for intermediate track).

In the NEGBP model, booklet easiness is explained by booklet properties. In our study we assembled booklets that explicitly differed in their a priori booklet easiness that was calculated as the mean of item easiness (from a pilot study) per booklet. This booklet property can be used as a predictor to check if the estimated booklet easiness is actually related to the a priori booklet easiness. Although we tried to hold any other booklet properties constant, this was not fully possible due to complex design restrictions. Especially, the number of items per booklet has a relevant variation with a range of 42 to 64 items ($M = 51.03$, $SD = 4.62$). This booklet property was centered and also added as a predictor.

In the *lme4* package function *glmer* with the argument *family* set to binomial (link = “logit”) is used for model estimation. This specifies the logit function as the link function that maps the location of the distribution of responses onto the latent scale, and sets this distribution as being a binomial distribution which corresponds to the Bernoulli distribution in the case of dichotomous responses. To indicate random effects the character | needs to be used. The syntax of our four models is:

```
Rasch (RPRI): Y ~ 1 + ( 1 | person ) + ( 1 | item ) ,
Booklet model: Y ~ 1 + ( 1 | person ) + ( 1 | item ) + ( 1 | booklet ) ,
NEGB model:   Y ~ 1 + ( 1 | person ) + ( 1 | item ) +
               ( 1 | booklet ) + track + ( 1 | booklet:track ) ,
NEGBP model:  Y ~ 1 + ( 1 | person ) + ( 1 | item ) +
               ( 1 | booklet ) + track + ( 1 | booklet:track ) +
               AprioriBookletEasiness + N_Item.
```

Results

Table 2 shows the estimated parameters of all four models. The Rasch model was estimated as a reference to gauge the fit of the other models. In the booklet model, booklet was added as a random effect. In a study design with completely randomly assigned booklets this effect of $SD = 0.34$ logits could be interpreted as the deviation of booklet easiness controlled for the easiness of items. However, in our study design, the distribution of booklets depended on the school track. Thus, the variable school track must be taken into account as a covariate. In the NEGB model, the fixed effect of school track is estimated as $\delta_{track} = 0.53$ logits. Since effect coding was used and there are two track groups, the difference in average competence between students attending the academic track and students attending the intermediate track is $2 * 0.53 = 1.06$ logits. The estimated booklet effect in this model is $SD = 0.08$ logits with a 95% confidence interval (CI), lower limit $LL = 0.05$, and upper limit $UL = 0.12$,

Table 2. Fixed and Random Effects Estimates and Model Fit of Rasch, Booklet, NEGB, and NEGBP model.

	Rasch model		Booklet model		NEGB model		NEGBP model	
	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI
Fixed effects								
Intercept	0.20	[0.13, 0.32]	0.17	[0.02, 0.30]	0.30	[0.24, 0.38]	0.31	[0.24, 0.38]
Group (school track)					0.53	[0.51, 0.54]	0.54	[0.52, 0.55]
A priori booklet easiness							0.14	[0.08, 0.21]
N Item							-0.014	[-0.020, -0.011]
Random effects								
Student	0.95	[0.94, 0.97]	0.90	[0.89, 0.92]	0.81	[0.80, 0.83]	0.81	[0.80, 0.83]
Item	1.33	[1.30, 1.39]	1.34	[1.30, 1.40]	1.35	[1.28, 1.41]	1.35	[1.28, 1.41]
Booklet			0.34	[0.29, 0.39]	0.08	[0.05, 0.12]	0.07	[0.03, 0.11]
Booklet × Group					0.05	[0.01, 0.12]	<0.001	[-0.068, 0.020]
Model fit								
AIC	550,203		549,369		547,670		547,648	
BIC	550,236		549,413		547,737		547,738	
Deviance	550,197		549,361		547,658		547,632	

Note. CI = confidence interval; N Item = number of items in a particular booklet; NEGB = nonequivalent groups booklet; NEGBP = nonequivalent groups booklet properties. For random effects the estimate reported is SD.

indicating a significant difference from zero. Thus, the used booklets vary significantly in their easiness (after controlling for item easiness). Although this effect is rather small (about 10% compared with the between-person variation), it still contributes in a nonnegligible manner to the probability to answer the test items correctly. The booklet $\times$ group interaction of $SD = 0.05$ logits (95% CI [0.01, 0.12]) indicates that booklet effects additionally vary between groups to a small degree. Thus, a specific booklet may be a little easier (or a little more difficult) for one school track compared with the other school track. The intercept in the NEGB model ($\alpha_0 = 0.30$) is slightly higher than in the Rasch model ($\alpha_0 = 0.20$). This is due to the nonuniform distribution of school track in our data and the uncentered effect coding ($-1/1$) of the track variable. While the Rasch intercept is representing the predicted mean for a population with 58.77% intermediate track and 41.23% academic track students, the intercept in the models that include school track as a predictor (NEGB/NEGBP) is the predicted mean for a population with 50% in each track. Thus, less competent intermediate track students are slightly underrepresented; therefore, the intercept is a little higher. This is just an effect of the coding and has no additional practical implications besides the interpretability of the intercept. Concerning model fit, the NEGB model performs much better than the Rasch model ($AIC_{diff} = 2,533$; $BIC_{diff} = 2,499$) giving additional evidence that the Rasch model is not the best choice to explain the response variation in the data at hand. Taking a closer look at the booklet estimates (calculated as best linear unbiased predictors; Bates, 2010) from the NEGB model reveals a range of booklet easiness from -0.12 to 0.13 logits, a result that is, for instance, in line with booklet difficulties reported by PISA (OECD, 2012) of -0.19 to 0.11 ($SD = 0.10$) for science booklets. To illustrate the impact of this booklet effect, we consider a simple example using estimates from the NEGB model. We assume a hypothetical person from an academic track school ($\delta = 0.53$) with a competence of $\theta = 0$ who solves a hypothetical item with an easiness of $\beta = 0$, while the model intercept is $\alpha_0 = 0.30$. If the item is in the easiest booklet, the correct response probability is $P(Y_{jibs} = 1) = \text{logit}^{-1}(0.30 + 0 + 0 + 0.53 + 0.13) = 0.723$, while the probability would be $P(Y_{jibs} = 1) = \text{logit}^{-1}(0.30 + 0 + 0 + 0.53 - 0.12) = 0.670$, if the same item is contained in the hardest booklet. Thus, solving this item is $0.723 - 0.670 = 5.3\%$ more likely if it would be deferred from the hardest to the easiest booklet.

While the NEGB model showed that a significant booklet effect is present in the data, the question why this effect occurs can be explored with the NEGBP model that incorporates booklet properties as predictors. The two booklet properties investigated in the present study are the a priori booklet easiness and the number of items per booklet. The a priori booklet easiness is significantly related to the estimated booklet easiness ($b = 0.14$, 95% CI [0.08, 0.21]), indicating that the specific compilation of items has an effect beyond the mere effect of item easiness. A booklet will get 0.14 logits easier if the a priori booklet easiness is increased by 1 logit. The negative coefficient of the number of items per booklet variable ($b = -0.014$, $[-0.020, -0.011]$) indicates that the higher the number of items in a booklet the more difficult the

booklet. For instance, a booklet becomes 0.14 logits more difficult if 10 more items are added. The log-likelihood ratio test comparing the deviance values of the NEGB and the NEGBP model is significant ($\chi^2_{\text{diff}} = 25.93$; $df_{\text{diff}} = 2$; $p < .001$) indicating that booklet properties can substantially explain the variation in booklet easiness.

Discussion

The objectives of this research are to introduce models that include booklets as a source of response variation in large-scale assessments and to illustrate how these models can be fitted and interpreted in a typical large-scale assessment setting. The inclusion of booklets in the psychometric analyses is important because the commonly applied IRT models assume that student responses are not affected by the assigned booklet. In fact, in conventional test settings with just one measurement instrument (booklet), there is no need to assume an empirical booklet effect since the measurement instrument is held constant. The necessity to be cautious about potential booklet effects recently began with the emergence of large item pools that required new approaches like multiple matrix sampling designs. In such designs the entire item pool is divided into smaller subsets (booklets) to keep assessment time for each individual student acceptable. Nevertheless, this approach comes with the price of potential booklet effects that can lead to biased results if not adequately accounted for. One reasonable strategy to avoid booklet effects is to keep all booklet properties—especially booklet easiness—constant. Still, this approach is dependent on prior knowledge of either booklet easiness or—as a proxy—of the mean of item easiness per booklet and thus is not feasible if these data are lacking. Furthermore, distributing booklets with varying easiness to subpopulations is not possible if booklet easiness is held constant.

The arguments to adjust the easiness of tests to students' competence levels are borrowed from the computerized adaptive testing literature, but they are relevant in paper-and-pencil settings as well. First, technical measurement properties (e.g., measurement precision) are increased. Second, motivational issues (e.g., boredom, frustration) may be ruled out. In a paper-and-pencil context, further research is needed to explore the magnitude of the assumed effects for both of these issues. Although we argue that competence-matched booklets are advantageous, we must acknowledge its limitations in the paper-and-pencil context: the easiness of a booklet cannot completely be matched to the competence of an individual student since the competence is unknown beforehand and usually varies notably between students. Therefore, each student receives a booklet that only approximately matches his or her competence. So even if booklets are matched to the assumed average competence levels of the subpopulations, there is usually still a mismatch on the individual level. However, this mismatch is on average not as large as in completely randomized distributions of booklets. Thus, while this approach is a step toward tailoring the test easiness to the competence level, full adaptation—as in computerized adaptive testing—surely cannot be achieved.

We used empirical data from a typical large-scale assessment study to illustrate the usage and interpretation of the proposed models and to investigate the magnitude of the empirical booklet effect. The estimated booklet effect in the booklet model was $SD = 0.34$ logits. In a study design with completely randomly assigned booklets, this effect would be interpretable as the dispersion of booklet easiness controlled for item easiness. However, in the study design of our empirical example, the distribution of booklets depended on two groups with different mean competences (i.e., nonequivalent groups) that received tailored booklets with varying easiness. The nonequivalent groups variable was school track with intermediate school track students' competence being 1.06 logits worse than academic track students' competence. When taking the nonequivalent groups into account in the NEGB model, the empirical booklet effect decreased to $SD = 0.08$ logits. Although this effect looks rather small at first sight, it is nevertheless nearly 10% compared with the random student effect ($SD = 0.81$). This result underlines the necessity to model the dependency of booklets and subgroups to attain unbiased booklet parameters. Furthermore, not accounting for this booklet effect (e.g., when using the Rasch model in a multiple matrix sampling design) may lead to either or both biased student competence estimates and biased item easiness estimates. Too broad or too narrow student distribution estimates cast a biased light on the competence of students, especially if mapped to levels of a predefined competence model that is used as a basis to compare students of different populations (e.g., Tiffin-Richards, Pant, & Köller, 2013). Biased item parameters are especially problematic if linking procedures are used to compare different populations. Unmodeled booklet effects that vary in their magnitude across populations might then inflate or deflate population differences in student competence.

Of course we must acknowledge that the effects found in our study may not generalize to other samples or test designs. Since only a few studies (primarily PISA) report booklet effects, there is a need for more research to investigate if booklet effects actually are a substantial phenomenon in large-scale assessment. Additionally interesting is the question *why* booklet effects occur to derive insights for the design process. Booklet properties that exert no effects on the responses can be ignored. Contrarily, booklet properties that influence test taking should be already handled in the design. We identified number of items as a relevant property that leads to more difficult booklets if increased. To avoid this effect, each booklet should contain the same number of items. In PISA (OECD, 2005), different locations of domains within each of the booklets is proposed as a potential cause of booklet effects, although this is not investigated. Further research should identify more relevant booklet properties to optimize designs in large-scale assessments.

Another limitation concerns the unbiasedness of our results. Since the true effects are unknown, one cannot be completely sure that even the assumingly correct NEGB model estimates are accurate. Simulation studies would help evaluate the magnitude of bias under certain conditions and test designs. In fact, there has been intensive research on a variety of specific issues concerning similar designs for nonequivalent

groups (see book chapters and cited references in Dorans, Pommerich, & Holland, 2007; Downing & Haladyna, 2006; Kolen & Brennan, 2004; von Davier, 2011), although these were not applied to the specific application of modeling booklet effects.

To conclude, assembling booklets with different average item easiness and distributing them to competence-matched subsamples is a worthwhile endeavor. However, practitioners should be aware that booklets may—intentionally or unintentionally—differ in their properties, and that these may have an effect on the responses given and hence on the parameter estimates of commonly used IRT models. As usually not all potentially effectual booklet properties are known—or if known cannot be held constant across booklets—it is advised to at least pay heed to the possible occurrence of booklet effects. A suitable approach is to use a statistical model to decide whether their magnitude is small enough to be ignored or if they should be accounted for in the subsequent analyses by explicitly including them in the scaling model. PISA (OECD, 2002, 2005, 2009, 2012) also proposed correction approaches for the case that adjustments of competence estimates for booklet effects are deemed necessary.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Institute for Educational Quality Improvement at Humboldt-Universität zu Berlin, Berlin, Germany.

References

- Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, 55, 92-104.
- Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Retrieved from <http://lme4.r-forge.r-project.org/book/>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-6)*. Retrieved from <http://CRAN.R-project.org/package=lme4>
- Beaton, A. E. (1988). *The NAEP 1985-86 reading anomaly: A technical report*. Princeton, NJ: Educational Testing Service.
- Beaton, A. E., & Zwick, R. (1990). *The effect of changes in the national assessment: Disentangling the NAEP 1985-86 reading anomaly* (Report No. ETS-17-TR-21). Princeton, NJ: Educational Testing Service.
- Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, 33, 609-624.
- De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73, 533-559.

- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, 39(12), 1–28.
- De Boeck, P., & Wilson, M. (2004). A framework for item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 3-41). New York, NY: Springer.
- Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. New York, NY: Springer.
- Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: Erlbaum.
- Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, 17, 345-356.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, 10, 133-170.
- Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, 28, 39-53.
- Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 3, 125-156.
- Hecht, M., Roppelt, A., & Siegle, T. (2013). Testdesign und Auswertung des Ländervergleichs [Test design and analysis of IQB national assessment]. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Eds.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* [The IQB National Assessment study 2012—Competencies in mathematics and the sciences at the end of secondary level] (pp. 391-402). Münster, Germany: Waxmann.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.
- Kremer, K., Fischer, H. E., Kauertz, A., Mayer, J., Sumfleth, E., & Walpuski, M. (2012). Assessment of standard-based learning outcomes in science education: Perspectives from the German project ESNAS. In S. Bernholt, K. Neumann, & P. Nentwig (Eds.), *Making it tangible: Learning outcomes in science education* (pp. 201-218). Münster, Germany: Waxmann.
- 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.
- Leucht, M., Harsch, C., Pant, H. A., & Köller, O. (2012). Steuerung zukünftiger Aufgabenentwicklung durch Vorhersage der Schwierigkeiten eines Tests für die erste Fremdsprache Englisch durch Dutch Grid Merkmale [Guiding future item development via predicting English as a foreign language item difficulties by Dutch Grid characteristics]. *Diagnostica*, 58, 31-44.
- Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come* (MESA Memorandum No. 69). Chicago, IL: University of Chicago, MESA Psychometric Laboratory. Retrieved from <http://www.rasch.org/memo69.pdf>

- Lord, F. M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement*, 22, 259-267.
- Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to educational standards: The impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education*, 8, 545-563.
- OECD. (2002). *PISA 2000 technical report*. Paris, France: Author.
- OECD. (2005). *PISA 2003 technical report*. Paris, France: Author.
- OECD. (2009). *PISA 2006 technical report*. Paris, France: Author.
- OECD. (2012). *PISA 2009 technical report*. Paris, France: Author.
- R Core Team. (2014). *R: A language and environment for statistical computing* (Version 3.1.0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Rutkowski, L., Gonzales, E., von Davier, M., & Zhou, Y. (2014). Assessment design for international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 75-95). Boca Raton, FL: CRC Press.
- Shoemaker, D. M. (1971). *Principles and procedures of multiple matrix sampling* (Report No. SWRL-TR-34). Inglewood, CA: Southwest Regional Educational Lab.
- Tiffin-Richards, S. P., Pant, A. H., & Köller, O. (2013). Setting standards for English foreign language assessment: Methodology, validation, and a degree of arbitrariness. *Educational Measurement: Issues and Practice*, 32, 15-25.
- van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York, NY: Springer.
- van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement*, 28, 317-331.
- von Davier, A. A. (Ed.). (2011). *Statistical models for test equating, scaling, and linking*. New York, NY: Springer.
- Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Erlbaum.
- Weirich, S., Haag, N., & Roppelt, A. (2012). Testdesign und Auswertung des Ländervergleichs: Technische Grundlagen [Test design and analysis of IQB National Assessment: Technical fundamentals]. In P. Stanat, H. A. Pant, K. Böhme, & D. Richter (Eds.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011* [Students' competences in German and Mathematics at the end of Grade 4: Results of the IQB National Assessment study 2011] (pp. 277-290). Münster, Germany: Waxmann.
- Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43-74). New York, NY: Springer.
- Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, 2, 1-17.
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10, 1-17.
- Wolf, L. F., & Smith, J. K. (2005). The consequence of consequence: motivation, anxiety, and test performance. *Applied Measurement in Education*, 8, 227-242.