Effects of Design Properties on Parameter Estimation in Large-Scale Assessments

Educational and Psychological
Measurement

2015, Vol. 75(6) 1021–1044

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164415573311

epm.sagepub.com

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

Abstract

The selection of an appropriate booklet design is an important element of large-scale assessments of student achievement. Two design properties that are typically optimized are the *balance* with respect to the positions the items are presented and with respect to the mutual occurrence of pairs of items in the same booklet. The purpose of this study is to investigate the effects of these two design properties on bias and root mean square error of item parameter estimates from the Rasch model. First, position effects are estimated using data from a large-scale assessment study measuring the competencies of 19,107 ninth graders in science. These results were then used for a simulation study with 1,540 booklet designs with systematically varied position balance and cluster pair balance. The simulation results showed a small effect of position balancing on bias and root mean square error of the item parameter estimates while the cluster pair balance was ignorable. This null effect is actually good news for test designers since it allows for deliberately reducing the degree of cluster pair balance without negative effects on item parameter estimates. However, it is recommended to try to achieve a high position balance when designing large-scale assessment studies.

Keywords

large-scale assessment, multiple matrix sampling, incomplete block designs, position effects, balancing, generalized linear mixed models (GLMM)

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

The goal of large-scale assessments of student achievement is to measure the competence of students in one or several domains. The focus of these assessments lies on deriving estimates for student populations or subpopulations and not on the level of individual students. The tests typically used to meet this purpose are assembled from a pool of items. Such an assembly scheme is called *test design* in general or *booklet design* in the case of large-scale assessments of student achievement. A popular approach to create booklet designs is multiple matrix sampling (Frey, Hartig, & Rupp, 2009; Gonzalez & Rutkowski, 2010; Rutkowski, Gonzales, von Davier, & Zhou, 2014; Shoemaker, 1971). The central idea behind this technique is to treat each student with only a subset of all items (usually called *booklet* in the case of paper and pencil tests, or more generally *test form*). Thereby, individual test burden can be kept in acceptable boundaries even though a large number of items can be incorporated in one study. However, besides these advantages, the construction of an appropriate booklet design is a nontrivial task. The question that instantly arises is how to assign items to booklets in an optimal way. It is easily imaginable that there typically are a vast amount of different options. Since a lot of options also imply lots of potential defects, it is of great importance to carefully choose the booklet design, as Giesbrecht and Gumpertz (2004) emphasize in their textbook on experimental design:

Although proper examination of the results of an experiment is important, there is no way that a clever analysis can make up for a poorly designed study, a study that leaves out key factors, or inadvertently confounds and/or masks relevant factors. (p. 2)

So what is the best booklet design? Unfortunately, there is no one best overall booklet design; instead, a booklet design needs always to be tailored to the specific objectives and has to take the restrictions of the study into account. However, a general statistical objective guiding the development of booklet designs is to strive for the unbiased estimation of the parameter estimates of interest (Frey, Hartig, & Rupp, 2009). A parameter estimate is unbiased if its expected value equals the true value of the parameter. If this is not the case, the estimate is biased. Bias can occur because of several sources. A common source is violations of assumptions of the applied statistical procedure. However, a difference between the parameter estimate and the true value of the parameter may also occur by model misspecifications, for instance if a relevant factor is not included. Such kind of bias has been called *omitted variable bias* (Greene, 2011; Wooldridge, 2013). In our article, we also subsume such misspecification effects under the term *bias*.

How can designs help to improve the accuracy of parameter estimates? One general approach is to *balance* certain design factors that are suspected to bias parameter estimates. Two factors that are commonly balanced in large-scale assessments are *item positions* and *item pairs*. Balancing item positions means to place each item at every position in the test with equal frequency while item pair balance is achieved by combining each item with all other items with equal frequency over the complete set of booklets. Although such balancing is popular in large-scale assessment

programs (e.g., in PISA, Organisation for Economic Co-operation and Development, 2012; in NAEP, Allen, Donoghue, & Schoeps, 2001; in German Science Education Standards (GSES), Hecht, Roppelt, & Siegle, 2013), the actual benefit has not been thoroughly investigated. If balancing leads to a huge decrease in bias, it is pivotal to balance designs. Otherwise, if nothing is gained by balancing, it might be ignored in designs for future studies. Hence, the general goal of this research is to explore the effect of balancing on the unbiasedness of item parameter estimates in the widely applied Rasch model. The results of the study will help practitioners to decide to what degree balancing is needed when using the Rasch model in large-scale assessment contexts.

The article is structured as follows: First, technical key terms of booklet designs are described. Second, the need for balancing booklet designs is discussed that will imply the research scope of this article. After that we will present a psychometric model that incorporates position effects and describe its application to data from a large-scale assessment of student achievement in science. This is done to derive a realistic data setting for a subsequent simulation in which design balance is systematically manipulated. The article ends with a discussion of the simulation results and concluding remarks concerning practical implications.

Booklet Designs

The basal units in booklet designs are items. However, items are grouped to disjunctive *clusters* of equal processing time to keep the design process manageable. Then, instead of items, balancing has to be conducted on the level of clusters. Booklet designs in which cluster positions and cluster pairs are balanced have been called *balanced incomplete block designs* in the literature (Giesbrecht & Gumpertz, 2004). These designs are insofar *incomplete* as each booklet contains only a subset of—instead of all—available clusters. This is practical since the number of booklets gets very fast unmanageably large in *complete* block designs.

We will adapt the terms and notation used in Giesbrecht and Gumpertz (2004) and Frey, Hartig, and Rupp (2009) with the slight modification of denoting absolute numbers with capital instead of small letters. Hence, the corresponding small letters can now be used as running indices. When building booklet designs, the most basic units are the items i . In a first step, the items are grouped to clusters t . The rationale behind creating clusters might differ from study to study, though one of the main advantages is that clusters of equal processing time (e.g., 20 minutes) can be assembled from items that usually differ in processing time. Clusters of equal length then remarkably facilitate the creation of booklets. After clusters have been defined, the total number R of repetitions of clusters must be chosen. In general, a higher number of repetitions increases the flexibility in the design since the number of possible cluster combinations rises but also implies a higher number B of booklets. We will call repetitions of clusters in different booklets *cluster instances* and denote them t_r with $r = 1, \dots, R$. The instrument that is given to individual students is a booklet. Each and every

booklet contains a fixed number of K clusters. Since students process the clusters one after another, there is an ordering of clusters that we call *cluster positions* p ranging from 1 to P (with $P = K$). The relation between the basic factors in booklet designs, items, clusters, booklets, and cluster positions, can be described with the terms *nested* and (completely or partially) *crossed* that are commonly used in the experimental design literature. A factor is nested within another factor if each level of the first factor co-occurs with only one level of the second factor. Two factors are completely crossed if every level of a factor co-occurs with every level of the other factor. If only some levels of two factors occur together, these factors are partially crossed. In booklet designs, the following relations of design factors are usually prevailing: items are nested within clusters; clusters are either nested in booklets (if clusters are repeated just once, $r = 1$) or crossed with booklets (if clusters are repeated more than once, $r > 1$, and the resulting cluster instances are assigned to several booklets). The crossing of clusters and booklets is complete in *Latin square designs* (e.g., Giesbrecht & Gumpertz, 2004). However, this complete crossing drastically increases the number and length of booklets with increasing number of clusters since $B = T = K$ in Latin square designs. Thus, in large-scale assessments of student achievement incomplete block designs (Frey, Hartig, & Rupp, 2009) with a partial crossing of clusters and booklets are usually used. Booklets and cluster positions are completely crossed, since every booklet contains all positions. Clusters might be nested in positions (each cluster appears at only one position); otherwise these two factors may also be partially crossed (each cluster appears at two or more—but not at all—positions), or completely crossed (each cluster appears at all positions).

Closely related to these concepts on the relation of design factors is the term *balance*. Unfortunately, definitions and usage of this term differ quite a bit in the design literature (see Preece, 1982, for a discussion) although there is agreement that absence of balance can lead to erroneous statistical results. We define balance as the equal occurrence of factor levels or combinations of factor levels. In this sense, a design is balanced with respect to a certain factor if its levels occur the same number of times. Regarding the combination of two factors, balance is achieved if factors are completely crossed.

As mentioned above, two targets of balancing are common in large-scale assessments: cluster positions and cluster pairs. A design is position balanced if clusters and positions are completely crossed, that is, each cluster occurs at each position with equal frequency. We call such a design a *position balanced design*. A design in which all pairs of clusters occur with the same frequency is a *cluster pair balanced design*. A design combining both properties has been introduced by Youden (1937, 1940) and thus is called *Youden square design*. An example of a Youden square design with 31 clusters and six positions is shown in Table 1. In fact, we used exactly this design to obtain empirical data for the analysis of position effects.

Although balance has been originally defined as a dichotomous property (either a design is balanced or not), designs actually possess a continuous degree of balance, that is, unbalanced designs differ in how much they are not balanced. Hence, while

Table 1. A Youden Square Design with 31 Booklets, 6 Positions, and 31 Clusters.

Booklet	Position					
	1	2	3	4	5	6
1	1	2	3	4	5	6
2	7	8	9	5	10	11
3	12	13	14	10	6	15
4	16	17	18	6	11	19
5	2	20	21	11	15	22
6	8	23	24	15	19	1
7	13	3	25	19	22	7
8	17	9	26	22	1	12
9	20	14	27	1	7	16
10	23	18	28	7	12	2
11	3	21	29	12	16	8
12	9	24	30	16	2	13
13	14	25	31	2	8	17
14	18	26	4	8	13	20
15	21	27	5	13	17	23
16	24	28	10	17	20	3
17	25	29	6	20	23	9
18	26	30	11	23	3	14
19	27	31	15	3	9	18
20	28	4	19	9	14	21
21	29	5	22	14	18	24
22	30	10	1	18	21	25
23	31	6	7	21	24	26
24	4	11	12	24	25	27
25	5	15	16	25	26	28
26	10	19	2	26	27	29
27	6	22	8	27	28	30
28	11	1	13	28	29	31
29	15	7	17	29	30	4
30	19	12	20	30	31	5
31	22	16	23	31	4	10

Note. Cells contain a total of 186 cluster instances (6 instances of each of 31 clusters). All instances of Cluster 1 are black-rimmed to illustrate position balance of this cluster. Since each cluster occurs at all six positions the design is position balanced. Clusters 5 and 13 are shaded grey to give an example of a cluster pair. With each cluster pair occurring exactly once the design is cluster pair balanced.

the completely balanced design possesses the maximum amount of balance, there are other partially balanced designs that are not completely balanced but still retain a certain degree of balance. As an example, consider the design in Table 1 with 31 booklets and 6 positions. Cluster 1 occurs at Position 1 in Booklet 1, on Position 2 in Booklet 28, on Position 3 in Booklet 22, on Position 4 in Booklet 9, on Position 5 in Booklet 8, and on Position 6 in Booklet 6. Thus, Cluster 1 is completely position

balanced. As each of the other clusters occurs on all positions too, the design is completely position balanced. However, switching Cluster 1 with Cluster 2 in Booklet 1 would lead to a slight unbalance of these two clusters (Cluster 1 occurs twice at Position 2; Cluster 2 twice at Position 1). Thus, this design would not be completely balanced any more but slightly unbalanced. The degree of unbalance can further be increased by switching more clusters. A strongly unbalanced design is shown in Table 2. Here, Cluster 1 occurs solely at one position (Position 1). Thus, this cluster is heavily unbalanced. As most of the other clusters are unbalanced too, the entire design is very position unbalanced. The cluster pair balance is best investigated by visually displaying the clusters and their pairwise occurrence. Figure 1 displays both designs from Table 1 (figure at top) and Table 2 (figure at bottom). These figures were generated with the R package *eatDesign* (Hecht, 2014) that uses the R package *igraph* (Csardi & Nepusz, 2014). The dark grey circles depict the 31 clusters while the grey lines indicate which clusters mutually occur in booklets. In the completely cluster pair balanced Youden square design, each cluster is connected with each other cluster. However, in the partially cluster pair balanced design each cluster is just combined with a few of the other clusters.

Position Effects

A well-documented source of bias are effects caused by the position an item is presented in a booklet (e.g., Albano, 2013; Debeer & Janssen, 2013; Hahne, 2008; Hohensinn et al., 2008; Hohensinn, Kubinger, Reif, Schleicher, & Khorramdel, 2011; Weirich, Hecht, & Böhme, 2014). Since each student is treated with more than one item for economic reasons, an order of item presentation is inherent. Moreover, the position of an item within a booklet usually varies across booklets. This variation of item positions can potentially affect the probability of a correct response. Such a phenomenon is called *position effects* that may be interpreted from either the item or the person side. Considered from the item perspective, item parameters like the item difficulty estimate depend on the item position. For instance, an item might be interpreted as more difficult if placed toward the end of the test. Viewing the phenomenon of position effects from the person side, the individual competence estimate of a person might decline toward the end of the test. In this case, the estimated competence is higher at the beginning of the test than at the end. Independent of the interpretation, their occurrence is typically explained by fatigue, motivational aspects, or training effects. Students may get more and more exhausted or demotivated during the test and thus perform worse at the end compared with the beginning. Contrarily, students might perform better when getting more familiar with the kind of test material.

The magnitude of position effects is reported to be considerable and nonignorable. For instance, in Weirich et al. (2014, online appendix) significant nonlinear position effects were found with a range of 0.24 logits between the first and the last cluster position (while the deviations of persons and items were $SD = 1.01$ and $SD = 1.22$, respectively). Concretely, the effects (easiness) of Positions 2, 3, and 4 with respect

Table 2. A Partially Balanced Design With 31 Booklets, 6 Positions, and 31 Clusters.

Booklet	Position					
	1	2	3	4	5	6
1	13	15	16	25	19	31
2	1	5	7	27	11	3
3	10	30	29	2	14	9
4	1	27	7	5	11	3
5	13	19	15	16	20	31
6	10	29	9	22	14	12
7	18	26	23	8	25	4
8	1	5	27	21	11	3
9	26	18	23	8	4	25
10	10	28	29	2	14	12
11	6	22	24	8	23	17
12	13	15	26	25	4	31
13	18	21	23	26	4	8
14	13	19	30	16	20	31
15	1	18	24	7	8	21
16	28	6	24	22	12	17
17	18	21	26	23	4	25
18	18	24	8	23	21	17
19	5	21	27	3	11	7
20	6	28	24	22	12	17
21	19	30	15	16	20	9
22	19	30	15	9	20	16
23	30	29	9	2	20	14
24	6	28	24	22	12	17
25	13	16	30	19	20	31
26	10	9	29	2	14	12
27	1	5	27	7	11	3
28	10	6	2	22	28	17
29	13	26	15	25	4	31
30	1	5	27	7	11	3
31	10	6	28	2	29	14

Note. Cells contain a total of 186 cluster instances (6 instances of each of 31 clusters). All instances of Cluster 1 are black-rimmed to illustrate position unbalance of this cluster.

to Position 1 were estimated as $p_2 = -0.10$, $p_3 = -0.06$, $p_4 = -0.24$, which corresponds to centered position effects (difficulty) of $p_1 = -0.10$, $p_2 = 0.00$, $p_3 = -0.04$, and $p_4 = 0.14$. Consider a small example using these results: If an item is placed at all four positions (i.e., this item is balanced concerning positions) the item estimate is not affected as position effects are averaged out (i.e., they sum to 0). If another item with the same difficulty is, for instance, only placed at Position 4, then this item

Figure 1. Two designs with 31 clusters.
Note. The design at top is the completely cluster pair balanced Youden square design displayed in Table 1. The design at bottom is the partially balanced design displayed in Table 2. The dark grey circles depict the clusters; the grey lines indicate which clusters mutually occur in booklets.

estimate is biased by 0.14, that is, the item appears 0.14 logits more difficult as it actually is. Thus, when comparing the difficulties of these two items, the second item appears 0.14 logits more difficult than the first although both items possess the same true difficulty. This is a severe problem in analyses that rely on the correct ordering of items. For instance, when using the bookmark method (Mitzel, Lewis, Patz, & Green, 2001) to find cutoff scores in a standard setting procedure (e.g., Cizek & Bunch, 2007) items are ordered by increasing difficulty. Then cutoff scores between competence levels or between pass and fail are set depending on this ordering. Items that are incorrectly located can then bias the results of such an endeavor.

Another example concerns the same item that is used in two studies. Building on the previous example, we now assume that one and the same item is placed on all four positions in the first study and only at Position 4 in the second study. Then—under the assumption of equal position effects and equal person competence distributions across studies—this item's difficulty estimate is 0.14 logits higher in the second study than in the first study just because it was placed at other positions in the test. This is, for instance, problematic if linking methods based on common items (e.g., Dorans, Pommerich, & Holland, 2007) are used to compare students' competencies of two studies since between-study differences of item estimates because of position effects may bias the results of such linking procedures (Meyers, Miller, & Way, 2009).

Cluster Pair Effects

The effects of cluster pair balance on item parameter estimates have not yet been sufficiently investigated within the context and terminology of booklet designs. However, the occurrence or nonoccurrence of specific pairs of clusters renders down to the issue of data missingness or—in other words—sparseness of data matrices. In this sense, combining just a few clusters in the booklet design—which would represent a large unbalance of cluster pairs—will lead to high data sparseness whereas a cluster pair balanced design provides full item covariance information. A large variety of missing data techniques exist (for an overview, e.g., Enders, 2010; Graham, 2012; Schafer, 1997). In this article, we focus on methods and software that are popular in the large-scale assessment context. Particularly, we use the maximum likelihood approach with the EM algorithm (Bock & Aitkin, 1981; Dempster, Laird, & Rubin, 1977) that is incorporated in the software ConQuest (Wu, Adams, Wilson, & Haldane, 2007) and the R package *TAM* (Kiefer, Robitzsch, & Wu, 2014). Although there is research on various properties and performance of this method, the specific aspect of missingness of item covariance information implied by cluster pair unbalance has not been systematically investigated.

Research Scope

The goal of this research is to explore the effect of booklet design balancing on item parameter estimates when using a simple and widely applied item response theory (IRT) model, the Rasch model. Specifically, the bias and root mean square error

(RMSE) of Rasch item difficulty estimates dependent on the degree of the design's position and cluster pair balance is investigated. It is expected that higher position balance yields on average more accurate item parameter estimates. Here, the central assumption is that a high design balance will compensate for negative effects of using a too simple ("wrong") model (the Rasch model) on essentially more complex data (i.e., data that contain position effects). Furthermore, it is assumed that the degree of cluster pair balance does not influence the accuracy of parameter estimates. However, it is an open research question how low the cluster pair balance—and thus, how sparse the item covariance information—may be before negative effects and/or estimation problems occur. In our view, it seems reasonable to pursue this research question as fully or partially cluster pair balanced designs are popular in large-scale assessments of student achievement (e.g., in PISA, Organisation for Economic Co-operation and Development, 2012; in NAEP, Allen, Donoghue, & Schoeps, 2001; in GSES; Hecht et al., 2013). If nothing is gained by this design property, it might be ignored in future studies.

Empirical Estimation of Position Effects

Sample and Booklet Design

Data from a large-scale assessment study in Germany (*IQB-Ländervergleich 2012*; Pant et al., 2013) that measures the attainment of GSES was used to derive realistic position parameter estimates for the subsequent simulation. The full sample consisted of 44,584 ninth-grade students. Participation was mostly mandatory; students were neither rewarded nor graded. For more information on the sample see Siegle, Schroeders, and Roppelt (2013). In this study, five booklet designs were simultaneously administered to account for various demands (Hecht et al., 2013). The present study uses only data gathered with one of these designs. This design is a Youden square design with $I = 386$ science items, $T = 31$ clusters, $B = 31$ booklets, and $P = 6$ positions. The booklets were randomly distributed to a subsample of $J = 19,107$ students. The testing time for the science items was 2 hours, partitioned in two 1-hour sessions with a break of 15 minutes in-between.

Models

We will use the generalized linear mixed models framework (GLMM; De Boeck, 2008; De Boeck & Wilson, 2004) to extend the Rasch model to include parameters for position effects. The dependent variable in the Rasch model are the dichotomous (correct = 1 vs. incorrect = 0) responses, Y_{ji} , for $j = 1, \dots, J$ persons and $i = 1, \dots, I$ items. The central postulation is that the probability $P(Y_{ji} = 1)$ of person j to correctly response to item i depends on the competence of a person, θ_j , and an item parameter, β_i . If β_i enters the formula with a "+" sign, it is interpreted as *easiness*; if it is included with a "-" sign, as *difficulty*. This issue is relevant when using different software. For instance, the R package *lme4* (Bates, Mächler, Bolker, & Walker,

2014; R Core Team, 2014) uses the “+” parameterization while the R package *TAM* uses the “−” parameterization. One assumption of the Rasch model is that the responses Y_{ji} are Bernoulli-distributed with location $P(Y_{ji} = 1)$. To map these probabilities onto the latent continuous scale, the logit link is applied (e.g., De Boeck & Wilson, 2004). Further assumptions can be made about the distribution of person and item parameters. In the classic formulation of the Rasch model, the items are modeled as fixed effects—that is, the model contains one point estimate for each item—while the person parameters are specified as a random variable with a distribution, which is commonly assumed to be normal with a certain mean and variance. This kind of Rasch model is therefore called random person—fixed item (RPFI, De Boeck, 2008). Of course, all other combinations are also possible, random person—random item (RPRI), fixed person—random item (FPRI), fixed person—fixed item (FPFI). The choice of the Rasch model derivative depends on the unit of interest and on the inference that one aims for. If conclusions about specific units (e.g., competence of specific persons or difficulty of specific items) should be drawn, then a fixed effects perspective is needed. Contrarily, if units are seen as exchangeable, effects should be modeled as a random variable in order to use the most parsimonious model and to emphasize the different conceptual focus. Inferences are then drawn with regard to the population the units are stemming from and not for specific units. In our case of modeling position effects, the units of interest are the positions; thus, their parameters are modeled as fixed effects, while persons and items are assumed to be random variables with zero means and variances σ_θ^2 and σ_β^2 , respectively. Using the “+” parameterization for the item difficulty parameters, the mathematical formulation of the RPFI Rasch model is

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

Now, this equation can easily be extended to incorporate position effects: For each position $p = 2, \dots, P$ a parameter δ_p is added as fixed effect:

$$P(Y_{jip} = 1) = \text{logit}^{-1}(\alpha_0 + \theta_{jp} + \beta_i + \delta_p) \quad (2)$$

While the intercept α_0 in the RPFI Rasch model is interpretable as the difference between the means of persons and items, it is now the difference between the means of persons and items at the Reference Position 1. For instance, δ_2 is the effect of Position 2 compared with Position 1. Furthermore, it is plausible to assume that persons vary in their estimated competence over the course of the test, which might be caused by changing emotional (frustration/boredom) and/or motivational states. Thus, it is appropriate to specify not only one position-independent person parameter, θ_j , but position specific parameters, θ_{jp} . These parameters can be assumed to be multivariate normally distributed, $\theta_{jp} \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma})$, with a vector of zeros as locations and a variance-covariance matrix,

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

The position specific variances σ_p^2 on the main diagonal of this matrix indicate how much variation exists between persons at each position of the test. The covariances $\sigma_{p_1 p_2}$ ($p_1 \neq p_2$) indicate how much the competence estimates at two positions are associated.

Using the function *glmer* from the R package *lme4* with the *family* argument set to binomial (link = “logit”), the RPRI Rasch model is specified as

$$Y \sim 1 + (1 \mid \text{person}) + (1 \mid \text{item})$$

Instead of the commonly applied *reference coding* with one position being the reference for the effects of the other positions we will use *deviation coding* (e.g., Hutcheson & Sofroniou, 1999) to allow for a more intuitive interpretation of position effects. To that end, the *position* variable containing the indices for the six positions (in R such a variable is called a *factor*) first needs to be coded into five new variables according to the deviation coding scheme: On the first variable (*pos1*), Position 1 is coded as 1, Position 6 as -1 , and all other positions as 0; on the second variable (*pos2*), Position 2 is coded as 1, position 6 as -1 , and all other positions as 0; and so on for Positions 3, 4, and 5. For Position 6 no new variable is needed nor allowed since the first five variables already contain the information whether Position 6 is used or not. However, the effect of Position 6 can be calculated as the sum of the effects of Positions 1 to 5 multiplied with -1 . Through deviation coding the position effects are now interpretable as the deviation from the grand mean (i.e., the intercept). The syntax of the position effects model is

$$Y \sim 1 + \text{pos1} + \text{pos2} + \text{pos3} + \text{pos4} + \text{pos5} + \\ (-1 + \text{position} \mid \text{person}) + (1 \mid \text{item})$$

Results

Table 3 shows the difficulty estimates for the RPRI Rasch and the position effects model. For a better comparability with other and future studies that examine position effects we provide standardized position effects that are normed on a standard deviation of $SD = 1$ of the latent scale. Such standardization is reasonable since the effects in logistic models are represented in reference to a constant error variance; thus, parameter estimates might appear deflated in models with different explanatory strength (De Boeck, 2008). Over the first three positions, a clear decrease from 0.09 to -0.07 logits can be observed. Interpreted from the item side, this means that the difficulty of an item increases by $0.09 - (-0.07) = 0.16$ during the first half of the test. After 1 hour of

Table 3. Fixed and Random Effects Estimates and Model Fit of RPRI Rasch Model and Position Effects Model.

	RPRI Rasch model		Position effects model		Stand.	
	Est.	95% CI	Est.	95% CI		
Fixed effects	Intercept	−0.15	[−0.276, −0.023]	−0.15	[−0.275, −0.028]	
	Position 1			0.23	[0.221, 0.240]	0.094
	Position 2			0.13	[0.116, 0.135]	0.051
	Position 3			−0.17	[−0.178, −0.158]	−0.069
	Position 4			0.043	[0.034, 0.052]	0.018
	Position 5			−0.045	[−0.054, −0.035]	−0.018
Random effects	Position 6 ^a			−0.19	—	−0.076
	Persons					
	Position 1	1.03	[1.02, 1.04]			
	Position 2			0.99	[0.97, 1.01]	
	Position 3			1.05	[1.04, 1.08]	
	Position 4			1.08	[1.06, 1.10]	
	Position 5			1.10	[1.08, 1.12]	
	Position 6			1.13	[1.12, 1.15]	
	Items	1.21	[1.13, 1.30]	1.13	[1.12, 1.15]	
				1.23	[1.14, 1.31]	
Model characteristics	$R^2_{\text{GLMM}(c)}$	0.433		0.451		
	AIC	1,494,055		1,486,894		
	BIC	1,494,091		1,487,234		
	Deviance	1,494,049		1,486,838		

Note. RPRI = random person–random item; CI = confidence interval; Stand. = position effects standardized on SD = 1 of the latent scale. $R^2_{\text{GLMM}(c)}$ is computed as described in Nakagawa and Schielzeth (2013). For random effects, the estimate reported is the SD. For all statistics, at least two significant digits are displayed.

^aDue to the deviation coding, the effect of Position 6 was not estimated but can be calculated as −1 times the summed effects of Positions 1 to 5.

testing (i.e., after Position 3), there was a 15 minutes break. The assumption is that this break provides the opportunity for students to rejuvenate and thus that position parameters are higher (less difficult) again. Indeed, Position 4 is easier than Position 3 (0.02 vs. -0.07) but still considerably less easy than Position 1 (0.02 vs. 0.09). After Position 4, the difficulty starts to increase again until it reaches its maximum at Position 6 (-0.08). The nonoverlapping confidence intervals indicate that the differences of the position effects are significant. However, this might be because of high statistical power in our large sample. To investigate the “practical significance,” we computed the explained variance, $R^2_{\text{GLMM}(c)}$, as described in Nakagawa and Schielzeth (2013). The position effects model ($R^2_{\text{GLMM}(c)} = .451$) explains 1.8% more variance than the Rasch model ($R^2_{\text{GLMM}(c)} = .433$). Expressed in Cohen’s f^2 (Cohen, 1988) this is equal to an effect size of $f^2 = 0.033$, which falls in between the suggested range of a small effect ($0.02 \leq f^2 < 0.15$). The standard deviations of position specific random effects of persons increase steadily from 0.99 (Position 1) to 1.13 (Position 6) while their correlations range from .89 to .97 indicating that the rank order of persons is very stable. Comparing the Rasch model with the position effects model ($\text{AIC}_{\text{diff}} = 7,161$; $\text{BIC}_{\text{diff}} = 6,857$; $\chi^2_{\text{diff}} = 7,211$; $df_{\text{diff}} = 25$; $p < .001$) clearly reveals that the position effects model is more suitable to explain the data at hand.

Simulation Study

This section describes the conducted simulation and is organized as follows: First, the operationalizations of *position balance* and *cluster pair balance* are introduced. These two design properties are the independent variables of the simulation study. Second, the designs used in the simulation and the algorithms to create such designs that differ with respect to these two design properties are described. Third, the models to generate the data are specified. In the fourth subsection, we outline the calculation of the accuracy parameters (bias, RMSE) that are the dependent variables in our study.

Operationalization of Design Balance

The cluster pair balance was computed as the percentage of realized cluster pairs in relation to all possible cluster pairs ranging from 0 (unbalanced) to 100 (balanced). The position balance is the correlation of positions and clusters in the design matrix (see Frey, Hartig, & Rupp, 2009, for examples). If clusters and positions are completely crossed—that is, each cluster occurs at each position—this correlation is zero. If each position only contains one cluster and each cluster occurs only at one position, the correlation is one. Thus this correlation characterizes the *unbalance* concerning positions. To reverse the polarity of this parameter and scale it comparable to the used index for the cluster pair balance, the correlation was transformed to a scale with a minimum of 0 (unbalanced) and a maximum of 100 (balanced). The R package *eatDesign* provides an easy-to-use function to calculate position balance and cluster pair balance of a design.

Generating Designs

For the simulation, 1,540 designs were created that differed in cluster pair balance and position balance. Cluster pair balance ranged from 32 to 100 in steps of 2; position balance from 14 to 100 in steps of 2. Creating designs that are even more unbalanced was not possible for technical reasons. In our study, each booklet contains more than one cluster; thus, there are inherently more than zero realized cluster pairs. Concerning position balance, some clusters always had to be placed on more than one position what made completely unbalancing impossible.

The following procedure has been applied to create the designs: First, an algorithm was programmed that iteratively modified the cluster pair balance until the target cluster pair balance was reached. The starting point was the Youden square design from the *IQB-Ländervergleich 2012* represented as a matrix with 31 booklets in rows, 6 cluster positions in columns and 31 clusters (or rather $31 \times 6 = 186$ cluster instances) as cells (see Table 1). The algorithm started with a random selection of one cluster instance. After that, a second cluster instance from the same column (i.e., cluster position) was randomly selected with the restriction that no cluster will occur twice in a booklet after swapping cluster instances 1 and 2. Finally, the two selected cluster instances were swapped. The resulting new design was kept if the cluster pair balance was closer to the target cluster pair balance or else abandoned. This process was repeated until the target cluster pair balanced with a tolerance of 0.10 was reached. The result of this process was 35 designs that differed in cluster pair balance but still held a position balance of 100.

A second algorithm modified the position balance. This algorithm used two subroutines, one for balancing (if the position balance of the current design was below the target position balance) and one for unbalancing (if the position balance of the current design was above the target position balance). The balancing started with randomly selecting one cluster instance from all clusters that were unbalanced (i.e., clusters that occurred more than once at least at one position). Of this cluster, one instance that occurred more than once at a position was randomly selected. This cluster instance then was swapped with a randomly selected instance of a second cluster from the same row (booklet) with the restriction that the first cluster did not already occur at the position of the second cluster instance, thus consequently balancing the first cluster. The subroutine for unbalancing started with a random selection of one instance of a cluster that was not already completely unbalanced (i.e., a cluster whose instances occurred at least at two positions). Two instances of this cluster at different positions were randomly chosen. The second instance was swapped with the cluster instance (of another cluster) in the booklet that was at the position of the first cluster instance. Thus, the balance of the design was decreased.

Figure 2 shows the 1,540 simulated designs characterized by cluster pair balance (y-axis) and position balance (x-axis). The Youden square design is balanced with regard to both properties. It is located at the upper right end of the matrix and is indicated by a black square. Diamonds (top row) are cluster pair balanced designs; triangles (right column) are position balanced designs.

Figure 2. 1,540 generated designs differing in position balance (x-axis) and cluster pair balance (y-axis).

Note. Grey dots indicate designs that are not completely balanced. Diamonds (top row) indicate cluster pair balanced designs; triangles (right column) indicate position balanced designs. The square in the upper right corner represents the completely balanced Youden square design.

Generating Data

To recap, the goal of the current research is to investigate the accuracy of item parameter estimates from the Rasch model depending on the balance of the booklet design and the complexity of the data. Thus, two simulation models were used to generate data: the Rasch model and the position effects model. The data from the Rasch simulation model serves as a baseline reference as the Rasch model is the “right” model for the Rasch data. Contrarily, data generated by the position effects simulation model contains positions effects. In this case, the Rasch model is a too simple (“wrong”) model. However, this model-data incongruity might be compensated by the degree of balance of the booklet design.

The Rasch simulation model resembles the Rasch model in Equation 1 and the position effects simulation model resembles the position effects model in Equation 2, but since the *TAM* package was used to increase the speed of the simulation, the item parameterization had to be with a “–” sign. In both simulation models booklets were randomly distributed to students and normally distributed item difficulty parameters β_i ($I = 386$) with a deviation of $SD_{\beta_i} = 1.22$ were used as true item parameters

β_i^{true} in all conditions and replications. We chose a sample size of persons of $J = 2,604$ for two reasons: first, it is a realistic sample size (not too small, not too large) in large-scale assessments of student achievement; second, to keep distributing booklets to persons easy, a number of persons was chosen that was dividable by the number of booklets with zero remainder. Thus, each of the $B = 31$ booklets was distributed to $2,604/31 = 84$ persons. Furthermore, since each booklet was distributed to the same number of persons and each item occurred in an equal number (6) of booklets, the number of persons per item was constant within and across designs ($N/\text{item} = 504$). However, the number of items per person ranged from 69 to 80 ($Mdn = 75$, $SD = 1.93$) since the booklets contained a slightly different number of items. In the Rasch simulation model, $J = 2,604$ person parameters, θ_j , were drawn from a standard normal distribution. In the position effects, simulation model person parameters, θ_{jp} , for $J = 2,604$ persons and $P = 6$ positions per person were drawn from a multivariate normal distribution using function *rsmvnorm* from the R package *SimCorMultRes* (Touloumis, 2014) with the empirical fixed position effects $\delta_1, \dots, \delta_6$ as means and the variance-covariance matrix Σ from the position effects model. For each of the 1,540 previously created designs, $w = 50$ data sets (replications) in each simulation condition were generated.

Model and Outcomes

A Rasch model was estimated using the R package *TAM* to model the responses for each design, data condition, and replication, which sums up to a total of 154,000 Rasch models. Since the accuracy of item parameter estimation is of interest in our simulation, items needed to be modeled as fixed effects. Thus, the classic formulation of the Rasch model (RPFI) was used. We focused on two indices for evaluating the accuracy of the item parameter estimates: The bias describing the mean difference of the estimated item parameters, β_{iw} , and the true item parameter, β_i^{true} , for each item i :

$$\text{bias}_{\beta_i} = \frac{1}{W} \sum_w^W (\beta_{iw} - \beta_i^{\text{true}}) \quad (3)$$

and the RMSE that is the root of the averaged squared difference between the item parameter estimates and the corresponding true item parameter:

$$\text{RMSE}_{\beta_i} = \sqrt{\frac{1}{W} \sum_w^W (\beta_{iw} - \beta_i^{\text{true}})^2} \quad (4)$$

While the bias_{β_i} specifies the mean inaccuracy of an item parameter estimate, the RMSE_{β_i} additionally takes the variability of the estimate into account. To derive aggregated statistics on the level of designs, these two item statistics are averaged:

Table 4. Descriptives of $bias_{\beta}$ and $RMSE_{\beta}$ for Two Simulated Data Conditions.

Statistic	Data	M	SD	Minimum	Maximum
$bias_{\beta}$	Rasch	0.013	0.00063	0.012	0.016
	Position effects	0.074	0.02888	0.018	0.129
$RMSE_{\beta}$	Rasch	0.114	0.00080	0.112	0.118
	Position effects	0.145	0.01795	0.114	0.182

Note. $N = 1,540$ designs. For all statistics, at least two significant digits are displayed.

$$bias_{\beta} = \frac{1}{I} \sum_i^I |bias_{\beta_i}| \quad (5)$$

$$RMSE_{\beta} = \frac{1}{I} \sum_i^I RMSE_{\beta_i} \quad (6)$$

The index $bias_{\beta}$ describes the average bias of all items in a design. The $RMSE_{\beta}$ specifies the average root mean squared error of all items in a design.

Results

Descriptives of $bias_{\beta}$ and $RMSE_{\beta}$ dependent on the data condition are shown in Table 4. The mean $bias_{\beta}$ is almost six times larger for the position effects data than for the Rasch data (0.074 vs. 0.013), a clear effect implying that the Rasch model should be used with caution when position unbalanced designs are applied to data in which the occurrence of position effects is likely. The mean $RMSE_{\beta}$ is somewhat larger in the position effects data (0.145) compared with the Rasch data (0.114). Regarding the variation in $bias_{\beta}$ and $RMSE_{\beta}$ huge differences appeared: while there is virtually no variation in the Rasch data ($SD_{bias_{\beta}} = 0.001$; $SD_{RMSE_{\beta}} = 0.001$; values are rounded to three digits) the position effects data show relatively large variation ($SD_{bias_{\beta}} = 0.029$; $SD_{RMSE_{\beta}} = 0.018$). This implies that there are designs that are less prone to produce inaccurate item parameter estimates when position effects are present. Additional investigation also revealed that the parameter estimate of the person competence distribution (σ_{θ}^2) in the Rasch data was highly accurate, that is, there is virtually no bias ($M_{bias} = 0.005$, $SD_{bias} = 0.003$) and only marginal RMSE ($M_{RMSE} = 0.031$, $SD_{RMSE} = 0.003$).

To test the influence of the design balance on the accuracy of item estimates, regression analyses for each dependent measure and data condition were conducted with position balance and cluster pair balance being the predictors (Table 5). The variance in $bias_{\beta}$ is almost completely ($R^2 = 95.1\%$) related to the design balance in the position effects data. The regression coefficient of position balance, $b = -0.036$ ($p < .001$), shows that the bias increases by $b \times SD_{bias_{\beta}} = -0.036 \times 0.029 = -0.001$ (i.e., a bias reduction of 0.001) if the position balance is increased by 1.

Table 5. $bias_{\beta}$ and $RMSE_{\beta}$ Predicted by Design Properties in Regression Analyses.

Statistic	Data	Parameter	Est.	SE	<i>p</i>	R^2
$bias_{\beta}$	Rasch	Intercept	-0.1445			.002
		Position balance	0.0033	0.00343	.331	
		Cluster pair balance	0.0032	0.00318	.302	
	Position effects	Intercept	2.0811			.951
		Position balance	-0.0363	0.00076	<.001	
		Cluster pair balance	0.0016	0.00072	.019	
$RMSE_{\beta}$	Rasch	Intercept	0.9876			.143
		Position balance	0.0042	0.00310	.184	
		Cluster pair balance	-0.0147	0.00287	<.001	
	Position effects	Intercept	2.1476			.956
		Position balance	-0.0369	0.00069	<.001	
		Cluster pair balance	0.0007	0.00065	.291	

Note. $N = 1,540$ designs. Dependent variables (row "Statistic") were z standardized.

Comparing the design with the weakest position balance in our study ($pb = 14$) with the completely position balanced design ($pb = 100$) yields a bias difference of $(100 - 14) \times 0.001 = 0.086$. Additionally, the completely position balanced designs show indeed a much lower mean bias of $M_{bias_{\beta}} = 0.02$ compared with the $pb = 14$ designs ($M_{bias_{\beta}} = 0.12$). Thus, position balancing has a small, but nonnegligible bias reduction effect. Although the coefficient $b = 0.002$ for cluster pair balance is significant at $\alpha = .05$ ($p = .019$), the effect is too small (only 6% of the position balance effect) to have practical implications for most applications. Concerning the $RMSE_{\beta}$, the same picture emerged: balancing designs with respect to positions leads to more accurate item parameter estimates, whereas balancing cluster pairs has a null effect.

Discussion

Choosing the right booklet design is an important and often challenging task of large-scale assessments of student achievement. The criteria that should be included in the decision for a specific design are the goals of the assessment and the analysis strategy. Although aspects likely to influence the parameter estimates of interest may be handled in the statistical analyses after data acquisition, it is generally advisable to already take preventive actions when designing the test. One strategy is to balance the booklet design with respect to certain design factors. Two main targets of balancing techniques are the positions of clusters in different booklets and the mutual occurrence of cluster pairs within the same booklet. Thus, designs may vary in the balance of positions and the balance of cluster pairs. The purpose of this research was to examine the effect of these two design properties on the accuracy of item parameter estimates in the Rasch model. We began with estimating position effects based on empirical data to derive realistic values for our simulation. Similar to other research, we found clear evidence for position effects in our data. Although

standardized effects and effect sizes were rarely reported, we argue that this would help to compare findings of position effects across studies and facilitates the evaluation of practical relevance. In our sample of ninth-grade students who worked 2 hours on science items (with a 15 minutes break after 1 hour), standardized position effects parameterized as “easiness” range from 0.09 logits at Position 1 to -0.08 logits at Position 6. Expressed in Cohen’s effect size measure f^2 , positions exhibit a small effect ($f^2 = 0.033$) on the probability of a correct response. While fixed main position effects were typically investigated, the change of students’ parameter dispersion across positions has not been studied so far. In our model, we additionally allowed this deviation to vary over positions. Our analysis revealed that the spread of persons’ estimated competencies increased during the test. A possible explanation for this effect is the occurrence of changing emotional and motivational states during the test. High-achieving students might start to enjoy the test more and more when they experience that they will succeed—thus their performance increases. Contrarily, low achieving students might become more and more frustrated because the items are too difficult and start to underperform. One or both of these two mechanisms might then more and more inflate the competence deviation over the test.

For the simulation study, two data conditions were created: Rasch conform data as a baseline and a more realistic position effects scenario that incorporated the position effects found in the empirical data. As expected, analyzing Rasch conform data with the Rasch model results in very accurate item parameter estimates. Furthermore, this accuracy is independent of the position balance of the design. This result was expectable because the used data did not contain systematic position effects. However, if position effects are present in the analyzed data—which can be readily assumed for most empirical data sets resulting from large-scale assessments—balancing positions is of importance. Our results show that the most position balanced design possesses the lowest bias. The bias difference between the most position balanced and the least position balanced design was 0.09. This is a small effect, but when considering the normal range of item estimates from -3.00 to 3.00 , it might still make a difference for subsequent analyses if items are on average biased by an effect of this size. Additionally—as this is only the mean effect—the bias for particular items can be much larger.

An important and debatable issue is the nature of the assumed mechanism that underlies the two investigated effects in this study. The source of position effects is the interaction of persons with the measurement instrument. Depending on psychological phenomena (e.g., boredom, frustration, excitement, fatigue) that might occur during the test, the position of an item contributes to the probability to solve that item. While such position effects are well-documented in the literature, research on the effect of cluster pairs is lacking. In consequence, the potential source of this effect is more than vague as well. This lack of knowledge is further highlighted by the fact that reasons for balancing cluster pairs are usually not pointed out. We framed cluster pair balance and corresponding effects as a statistical problem and speculated that the statistical procedures might run into estimation problems if the available item

covariance information, which is determined by the cluster pair balance, becomes too sparse. Consequently, no data-generating simulation model for cluster pair effects is needed. Furthermore, the interpretation of the found null effect of cluster pair balance must be in accordance with the approach taken. This null effect indicates that the used statistical procedures are robust against sparse item covariance information. Other conclusions that are based on other assumptions on the nature of cluster pair effects cannot be drawn. Overall, more research on cluster pair effects is needed.

Several limitations need to be clearly stated. The results and implications do only apply for the specific, though typical, context that was studied, namely, item difficulty parameter estimates obtained in a RPFI Rasch model that was estimated using a marginal maximum likelihood approach in conjunction with an EM algorithm, and a linked incomplete block design with booklets randomly assigned to a medium-sized sample of students from a population with normally distributed competence. Although tempting, results cannot directly be generalized to other settings. For other parameters (e.g., person parameters), other statistics (e.g., Q3; Yen, 1984), other models (e.g., 2PL, 3PL), other estimation procedures (e.g., joint maximum likelihood, JML), other designs (e.g., designs with sets of unlinked booklets), other booklet distribution (e.g., when tailoring booklet difficulty to subsamples with differing competence), other software, and for other sample sizes, results concerning the relation of design balance and accuracy of parameter estimates may vary. Furthermore, these relations need not to be linear. Although visual inspection of our data did not give reason to drop the assumption of linearity, the range of design balance was trimmed at the left “unbalance” side because of the specific designs used (with 6 cluster positions and 31 clusters). Other designs may possess even markedly lower balance values than our least balanced designs studied. For these, bias might be higher than expected from a linear extrapolation of our effects.

To conclude, practitioners should be aware of position effects that can potentially bias item parameter estimates. However, the Rasch model can still be used despite the occurrence of position effects if the design is completely position balanced. For less position balanced designs it should be well considered if the bias is negligible enough. As a rule-of-thumb, increasing the balance of a design by 1 on the suggested position balance scale (ranging from 0 to 100) decreases the average item bias by 0.001. Another important finding is that balancing designs with respect to cluster pairs has no substantial effect on item parameter bias. This is actually good news for test designers since it allows for deliberately reducing item covariance information without negative effects on item parameter estimates. This opens up the possibility to use designs that are perfectly position balanced but only nearly cluster pair balanced. Designs of this kind are frequently produced if it is attempted to construct Youden square designs cyclically. Thus, the results of the present study are providing evidence that it is unproblematic to apply these designs, which is helpful because many large and very large Youden square designs are not yet found and cannot easily be constructed.

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

- Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement, 50*, 408-426.
- Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (No. NCES 2001-509). Washington, DC: National Center for Education Statistics. Retrieved from <http://nces.ed.gov/nationsreportcard/pdf/main1998/2001509.pdf>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-6) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=lme4>
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Cohen, J. E. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Csardi, G., & Nepusz, T. (2014). igraph: Network analysis and visualization (Version 0.7.1) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=igraph>
- De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533-559.
- De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.
- Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*, 164-185.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*, 1-38.
- Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales*. New York, NY: Springer.
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.
- 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*(3), 39-53.
- Giesbrecht, F. G., & Gumpertz, M. L. (2004). *Planning, construction, and statistical analysis of comparative experiments*. Hoboken, NJ: Wiley-Interscience.
- Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. In *IERI monograph series: Issues and methodologies in large-scale assessments (Vol. 3, pp. 125-156)*. Retrieved from <http://>

- www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_03_Chapter_6.pdf
- Graham, J. W. (2012). *Missing data: Analysis and design*. New York, NY: Springer.
- Greene, W. H. (2011). *Econometric analysis* (7th ed.). Harlow, England: Pearson Education.
- Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly*, 50, 379-390.
- Hecht, M. (2014). eatDesign (Version 0.0.10) [Computer software]. Retrieved from <http://R-Forge.R-project.org/projects/eat>
- 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.
- 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 Quarterly*, 50, 391-402.
- Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation*, 17, 497-509.
- Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. London, England: Sage.
- Kiefer, T., Robitzsch, A., & Wu, M. (2014). TAM: Test Analysis Modules (Version 1.0-2.1) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=TAM>
- 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.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.
- Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4, 133-142.
- Organisation for Economic Co-operation and Development. (2012). *PISA 2009 Technical Report*. Paris, France: Author.
- Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (2013). *The IQB national assessment study 2012: Competencies in mathematics and the sciences at the end of secondary level*. Münster, Germany: Waxmann. Retrieved from https://www.iqb.hu-berlin.de/laendervergleich/laendervergleich/lv2012/Bericht/IQB_NationalAssessment.pdf
- Preece, D. A. (1982). Balance and designs: Another terminological tangle. *Utilitas Mathematica*, 21C, 85-186.
- R Core Team. (2014). R: A language and environment for statistical computing. (Version 3.1.0) [Computer software]. 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.

- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London, England: Chapman & Hall.
- Shoemaker, D. M. (1971). *Principles and procedures of multiple matrix sampling* (Report No. SWRL-TR-34). Inglewood, CA: Southwest Regional Educational Lab.
- Siegle, T., Schroeders, U., & Roppelt, A. (2013). Anlage und Durchführung des Ländervergleichs [Concept and implementation 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. 101-121). Münster, Germany: Waxmann.
- Touloumis, A. (2014). SimCorMultRes: Simulates Correlated Multinomial Responses (Version 1.2) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=SimCorMultRes>
- Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement*, 38, 535-548.
- Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). Mason, OH: South-Western.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest version 2.0: Generalised item response modelling software [Computer software]. Camberwell, Victoria, Australia: ACER.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.
- Youden, W. J. (1937). Use of incomplete block replications in estimating tobacco-mosaic virus. *Contributions from Boyce Thompson Institute*, 9, 41-48.
- Youden, W. J. (1940). Experimental designs to increase accuracy of greenhouse studies. *Contributions from Boyce Thompson Institute*, 11, 219-228.