Psychological Test and Assessment Modeling, Volume 54, 2012 (4), 397-417

On the importance of using balanced booklet designs in PISA

Andreas Frey[1] *& Raphael Bernhardt*[2]

Abstract

The effect of using a balanced compared to an unbalanced booklet design on major PISA results was examined. The responses of 39,573 students who participated in the PISA-E 2006 assessment in Germany were re-analyzed. Using an unbalanced booklet design instead of the original booklet design led to an increase in mean reading performance of about six points on the PISA scale and altered the gender gap in reading to different degrees in the 16 federal states of Germany. For students with an immigration background, the reading performance was significantly higher for the unbalanced design than for the original design. For the unbalanced design, the relationship between self-reported effort while taking the test and reading performance was higher compared to the original design. The results underline the importance of using a balanced booklet design in PISA in order to avoid or minimize bias in population parameters estimates.

Key words: booklet design, testing, large-scale assessment, item response theory, Programme for International Student Assessment

[1] *Correspondence concerning this article should be addressed to:* Andreas Frey, PhD, Institute of Educational Science, Department of Research Methods in Education, Friedrich-Schiller-University Jena, Am Planetarium 4, D-07737 Jena, Germany. E-mail: andreas.frey@uni-jena.de

[2] Institute of Educational Science, Department of Research Methods in Education, Friedrich-Schiller-University Jena, Germany

Introduction

Large-scale assessments (LSAs) of student achievement aim to measure what populations of students know and can do in specified content areas. In LSAs large samples of students are assessed. Analyses of the observed responses make it possible to draw valid conclusions about the achievement levels in the underlying population of students. Many countries or federal states within countries run national LSAs. In the United States of America, for example, the *National Assessment of Educational Progress* (NAEP; e.g., Jones, & Olkin, 2004) has been conducted from 1969 on. The attainment of the *German national educational standards* is also assessed with LSA methodology (e.g., Stanat, Pant, Böhme, & Richter, 2012). In addition to the national initiatives, several international LSAs were initiated. One of the first was the *Pilot Twelve-Country Study* (Foshay, Thorndike, Hotyat, Pidgeon, & Walker, 1962) which was conducted in the year 1960. Some of the best known current international LSAs are the *Programme for International Student Assessment* (PISA; e.g., OECD 2010), the *Trends in International Mathematics and Science Study* (TIMSS; e.g., Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009), and the *Progress in International Reading Literacy Study* (PIRLS; e.g., Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009). The importance of national and international LSAs has increased steadily over the last decades. Today, these studies represent a core aspect of many educational systems around the globe.

Typically, LSAs strive to obtain reliable and valid information about student achievement in one or several content domains of interest. From a measurement point of view, the aspects of student achievement focussed on by LSAs are generally conceptualized as complex constructs. Often, differentiations in terms of subdimensions and/or facets are made. These subdimensions and/or facets represent aspects like cognitive processes, content areas, situations, or other systematizations of the respective content domain (e.g., OECD, 2009a for PISA 2009). Due to the complexity of the constructs at stake, large sets of items are required for their adequate operationalization. As an example, in PISA, about 150 to 200 items are employed to measure students' literacy in reading, mathematics, and science in each assessment. Such large numbers of items cannot be presented as a whole to each student within a realistic amount of testing time. Therefore, in LSAs, students are generally randomly given one of several test forms called booklets. Each booklet contains a subset of the complete item pool that can be sensibly answered by a student within a reasonable testing time.

The way the items are assigned to the booklets is specified by a booklet design (Frey, Hartig, & Rupp, 2009; Gonzales & Rutkowski, 2010; Yousfi & Böhme, 2012). In several LSAs like PISA or NAEP, a *Youden square Design* (YSD) is used as booklet design which is a special *balanced incomplete block design* (BIBD) (e.g., Giesbrecht & Gumpertz, 2004). From a statistical point of view, YSDs are well suited as booklet designs because they make it possible to control for two variables (in LSAs booklets and item or cluster position) which may have an unwanted impact on the parameter estimates of interest (Frey, Hartig et al., 2009).

YSDs were introduced in PISA 2003 after the variation of proficiency means between booklets was observed to be greater than expected in PISA 2000. These differences were

up to one quarter of a standard deviation (OECD, 2002). Further analyses indicated that the most likely explanation for the observed differences between the booklets were position effects. For most items (or more precisely: item clusters), the later it was presented in a booklet, the lower the observed relative frequency of a correct answer was. Fatigue, a reduction in test-taking motivation with increasing test length, or a lack of time on the part of the student at the end of the test may have produced this pattern. Furthermore, the results reported by OECD (2002) indicate that for some item clusters carry-over effects might have contributed to the differences between booklet means. To ensure a precise estimation of item parameters and, in turn, valid criterion-referenced interpretations in terms of proficiency levels, the differences between booklets were accounted for within the scaling procedure. This was done by applying a booklet correction for the estimation of item parameters by including the variable booklet into the scaling model (ConQuest statement: item + item*step + booklet). For the following estimation of the literacy distribution on the population level, the effect-coded variable booklet was used as conditioning variable in the background model using senate weights. Thus, the correction was carried out on the level of booklets. A more complex correction in terms of item positions in a booklet or in terms of cluster positions in a booklet would have produced a very complex model which would be problematic to handle in the international scaling (cf. OECD, 2002). As an additional reaction to the unexpectedly large differences between booklets, starting with PISA 2003, a YSD was used as booklet design in PISA. By presenting each item cluster in each of the four cluster positions exactly once in the set of 13 booklets, this YSD ensures that differences between booklet means will not affect country means. Note that even though the YSD is a clear improvement compared to the unbalanced booklet design applied in PISA 2000, position and carry-over effects are not removed but only averaged across positions. Consequently, a booklet correction was also applied from the PISA 2003 assessment on.

The booklet effects which are estimated in the scaling procedure can be interpreted as the amount that must be added or subtracted to the proficiency of a student who responded to one booklet in order to calculate the mean proficiency score, which is the score he or she would have achieved when answering all 13 booklets. Note that applying the booklet correction does not affect country proficiency means, but avoids potential problems when analyzing small subsamples where a uniform distribution of booklets might not be guaranteed due to small sample sizes. In PISA 2000, the largest observed booklet effect was 26.7 points on the PISA scale ($M = 500$; $SD = 100$) compared to the mean booklet difficulty. The mean absolute booklet effect was 11.7 points (OECD, 2002). In the following cycles, comparable booklet effects were observed. In PISA 2003, the maximum booklet effect was 37.2 points (average absolute booklet effect: 16.8 points; OECD, 2005), in PISA 2006, 40.4 points (average absolute booklet effect: 13.4 points; OECD, 2009b), and in PISA 2009, 31.1 points (average absolute booklet effect: 7.7 points; OECD, 2012).

Although the booklet effects reported for PISA seem to be of noteworthy magnitude, two important questions regarding booklet designs are still left unanswered by the results given in the PISA technical reports (OECD, 2002; 2005; 2009b). First, the results do not provide an answer to the question of which problems are avoided by using a balanced YSD instead of the unbalanced booklet design of PISA 2000. This would be of great

interest in order to learn more about the usefulness of YSDs in LSAs. If including booklet information in the scaling procedure, as done in PISA, is sufficient to account for position and carry-over effects, a YSD may not be necessary at all. From a practical point of view, this might be desirable for two reasons. The first reason is that YSDs restrict the test development process: YSDs only exist for some combinations of items, booklets, and positions within a booklet. Thus, within the test development process, an item pool needs to be constructed that fits into the structure of a pre-defined design and not vice versa. The second reason is that the administration of tests based on YSDs consumes a lot of resources and is more prone to scoring mistakes: YSDs typically comprise a medium to large number of different booklets. The YSD used in PISA, for example, has 13 booklets. If an unbalanced design had been used for presenting the same item pool, four booklets would have been sufficient. Substantial resources are needed for the type setting, formatting and handling of the different booklets.

A second question which is not answered by the results given in the PISA technical reports is the actual effect of using a YSD instead of an unbalanced booklet design on the final results given in the PISA reports. The available results which only quantify effects on the booklet level, do not make it possible to gauge whether relevant shifts of statistics have to be expected on the reporting metric when using an unbalanced design or not. In this regard, it is of upmost importance to check whether the assumption that booklet effects are the same for all students holds (this is what is controlled for by the booklet correction carried out in PISA) or whether effects differ between subgroups and thus affect the statistics of subgroups differently. If the latter is the case, the current procedure used in PISA for dealing with booklet effects may be problematic regarding the validity of the inferences drawn from statistics at the subpopulation level. Thus, it is necessary to know whether using different designs in which the position and carry-over effects are balanced to different degrees only has a main effect on PISA results or whether the effect differs between subgroups; which would mean that there is an interaction between the design and variables stratifying subpopulations. Interactions of this kind may lead to biased comparisons between subpopulations.

In order to provide answers to these two open questions, the current study examines the effect of using a YSD compared to an unbalanced booklet design on several key results of PISA. The rest of the text is organized as follows. First, the structure of the YSD design used in the PISA assessments from 2003 on is presented. Reasons are given for why this special booklet design structure was selected. Then, the research questions are stated. After that, the methods of the study and the results are described. The paper closes with a discussion on the usefulness of YSDs in PISA and in LSAs in general.

The PISA booklet design

The general purpose of booklet designs lies in distributing items to the participants in a way which fosters an unbiased and efficient estimation of the parameters of interest. What is of primary interest in the PISA main studies are country- and subgroup-specific means and variances for reading literacy, mathematical literary, and scientific literacy, as well as the co-variances between these domains. To derive these statistics, the mixed

coefficients multinomial logit model (MCMLM; Adams, Wilson, & Wang, 1997) is used for scaling the responses. The MCMLM is a generalized multidimensional Rasch model which can be combined with a population model, making it possible to estimate multidimensional distributions conditional on background variables (Adams, Wilson, & Wu, 1997). By incorporating background variables by means of a latent regression, not only the first and the second moment of the overarching multidimensional distribution for one country are estimated but also the moments of subgroup-specific distributions nested in the overarching distribution. This permits, for example, to report means and variances for students who have a different socio-economic background or who attend different school types within countries.

In order to estimate such complex multidimensional distributions, and to ensure that data gathering can be administered without problems, the PISA booklet design has to fulfill at least four requirements. First and obviously necessary, the single booklets have to be of comparable length to obtain similar answering times for all students. Second, the number of items per dimension needs to be controlled for on the sample level to achieve the desired content coverage. This is necessary because in each PISA assessment one of the three measured literacy domains serves as the major domain. For the major domain, more items are presented than for the other two domains, making it possible to report results for the subdimensions of the major domain. Third, all items measuring the same dimension should contribute an equal amount of information to the respective latent literacy dimension. This can be achieved by (a) presenting all items to the same number of students, and (b) by presenting each possible pair of items together equally often in a booklet. Fourth, the potential effect of the position of an item within a booklet on the probability of solving it needs to be controlled for. Position effects are present if the probability of answering an item correctly depends upon its position within a booklet. As described above, position effects have been observed in the PISA assessments. Similar results are reported by Zwick (1991) for NAEP, and Way, Carey, and Golub-Smith (1991) for the Test of English as a Foreign Language (TOEFL). Nevertheless, in the study of Hohensinn, Kubinger, Reif, Schleicher, and Khorramdel (2011) position effects of negligible size are reported for the mathematical competence test of the Austrian Educational Standards for 4th graders. A comprehensive list of papers covering position effects in LSAs can be found in Meyers, Miller, and Way (2009). Besides a number of statistical problems, the presence of position effects may jeopardize criterion-referenced test score interpretations. The reason for this lies in the fact that after the data collection has taken place, it cannot be determined whether an item was answered incorrectly because the student was in fact not proficient enough to solve it or just because of fatigue, a lack of test-taking motivation, or a lack of time. Hence, if position effects are present and not accounted for in an appropriate way, the proportions of students at low proficiency levels will be too high and the proportions at high proficiency levels too low. This would provide an incorrect picture of the proportions of students in the underlying population that know and can do certain things.

The YSD applied in PISA from the year 2003 on meets the four requirements mentioned above. To keep the booklet design manageable, single items are grouped to so-called item clusters before they are assigned to specific cells of the design. The general characteristics of a YSD transposed into the context of LSAs are:

1. Every booklet is of identical length, containing the same number of clusters.
2. Every item cluster appears equally often across all booklets.
3. Each combination of a pair of item clusters appearing together in the same booklet is the same for all possible pairs of clusters.
4. Every item cluster occurs at most once in a booklet.

The booklet design of PISA 2006 satisfies the four conditions mentioned above. It is shown in Table 1. The design comprises 13 booklets, each with four cluster positions. Thus, all booklets contain the same number of item clusters and – since the item clusters are composed to require a comparable amount of testing time – are of equal length. Furthermore, every item cluster appears in every position exactly once. Thereby, potential position effects are averaged on the level of item clusters. In PISA 2003, 2009, and 2012 a YSD with the same structure was used, but different item clusters were assigned to the 13×4 cells of the design.

Table 1:
Booklet Design of PISA 2006

	Booklet												
Position	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	S1	S2	S3	S4	S5	S6	S7	M1	M2	M3	M4	R1	R2
2	S2	S3	S4	M3	S6	R2	R1	M2	S1	M4	S5	M1	S7
3	S4	M3	M4	S5	S7	R1	M2	S2	S3	S6	R2	S1	M1
4	S7	R1	M1	M2	S3	S4	M4	S6	R2	S1	S2	S5	M3

Note. S = science, R = reading, M = mathematics, numbers indicate different cluster of a domain.

Research questions

The booklet design in Table 1 seems to meet the general requirements of PISA very well. But, as stated above, it is not known so far which problems are avoided by using this special kind of design, whether using a YSD may be obsolete when booklet information is considered in the scaling procedure, what potential effects are caused on the PISA results by using a YSD instead of an unbalanced design, and whether these potential effects differ between subgroups or not.

Summing up, the following three research questions are examined with the current study:

1. What effect does using a YSD compared to an unbalanced booklet design have on key PISA results?
2. Do the potential effects of using a YSD compared to an unbalanced booklet design on key PISA results differ between subgroups of the student population?
3. Which students are advantaged or disadvantaged by using a YSD instead of an unbalanced design?

Note that the paper focuses on the impact of using a balanced versus an unbalanced booklet design on key PISA results at the population level and thus on a high level of aggregation. It thereby contributes to a better understanding of the relevance of booklet and position effects for the validity of inferences drawn from PISA results. It explicitly does not focus on an in-depth analysis of position or booklet effects because these analyses typically will not provide answers regarding the relevance of test score interpretations at the population level.

Method

The three research questions were answered by re-analyzing the responses of the students who participated in the German PISA-E assessment in the year 2006 (Prenzel et al., 2008). The PISA-E 2006 assessment was conducted to allow comparisons to be made between the 16 federal states of Germany, which is not possible with the regular PISA sample drawn for the international comparison. PISA-E 2006 followed the same procedures and used the same item pool for the assessment of students' literacy in reading, mathematics, and science as the regular PISA 2006 assessment.

Sample

For the present study, the responses of $N = 39{,}573$ 15-year-old students were re-analyzed. Student sampling was based on a stratified sampling frame using the two-step procedure typically applied in PISA (e.g., Frey et al., 2008; OECD, 2009b). The sample makes it possible to draw inferences about the population of 15-year-old students in Germany when using appropriate student weights.

Design

To analyze the effect of the booklet design characteristics on the results of PISA-E 2006, three different booklet designs were specified. First, the *original* booklet design as used in PISA 2006 was used without any changes (Table 1).

Second, an *unbalanced* booklet design was specified by using only some booklets from the original design and disregarding the others. The booklets to be disregarded were chosen under the objective of maximizing the potential effects on the parameter estimates of interest while dropping as few booklets as possible. As mentioned above, substantial position effects can be expected in the PISA assessments with performance decreasing towards the end of the test. Hence, deleting booklets which have clusters for the same content domain in the last position should cause pronounced effects on the literacy estimates. Concerning the three literacy scales, the largest effects can be expected for the reading scale because it is covered with a smaller number of items in PISA 2006 than the other two scales. Due to the aforementioned considerations, booklets number two and number nine were deleted from the original booklet design. Both booklets have a reading

cluster in the last position. Note that the deletion of other booklets would have provided other unbalanced designs and other results. The structure of the resulting design is shown in Table 2. It is quite unbalanced: Item clusters are not presented equally often across the remaining 11 booklets. In fact, five item clusters occur four times, six item clusters three times, and one item cluster only two times. Additionally, pairs of item clusters occur together in a booklet with different frequencies across the 11 booklets. Lastly, only five item clusters are presented in all cluster positions.

As the third booklet design condition, the structure of the original booklet design from PISA 2006 was not altered but 2/13 of the responses were randomly deleted. This condition was introduced because statistics based on variances are directly related to the number of responses available for their calculation. Thus, with the condition *random deletion*, differences in statistics based on variances between the unbalanced design and the PISA booklet design can be directly compared.

Table 2:
Unbalanced Booklet Design

	Booklet										
Position	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	S1	S3	S4	S5	S6	S7	M1	M3	M4	R1	R2
2	S2	S4	M3	S6	R2	R1	M2	M4	S5	M1	S7
3	S4	M4	S5	S7	R1	M2	S2	S6	R2	S1	M1
4	S7	M1	M2	S3	S4	M4	S6	S1	S2	S5	M3

Note. S = science, R = reading, M = mathematics, numbers indicate different clusters of a domain.

Procedure

The original responses gathered in PISA-E 2006 were used for the additional analyses. In the condition *original* all available responses were used. For the conditions *unbalanced* and *random deletion*, the responses not covered by the respective design were deleted and treated as missing by design. Based on the reduced response sets, the same scaling procedures as had been used for the PISA-E 2006 study were accomplished separately for each condition. Detailed descriptions of these procedures can be found in OECD (2009b) and in Frey, Carstensen, Walter, Rönnebeck, and Gomolka (2008).

The software ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007) was used for scaling purposes and SPSS 19 for all other analyses. The scaling procedure can be subdivided into six steps. First, item difficulties were estimated for reading, mathematics, and science with a unidimensional Rasch model for each content domain. Booklet effects were included in the measurement model to prevent confounding item difficulties and booklet effects. Second, a background model was specified. Third, two five-dimensional Rasch models were estimated for each of the 16 federal states of Germany. The background model from the previous step was used for conditioning, including the variable

booklet (effect coded). Five plausible values (PVs) were drawn for each student. Fourth, the PVs were transformed to the PISA reporting metric. The linear expressions applied for the transformations are given in OECD (2009b). Fifth, the sample weights and replicate weights were recalculated. This was done to ensure that valid standard errors could be estimated for the two reduced data sets. Finally, the standard errors of the statistics of interest were estimated with the balanced repeated replication method using the SPSS Replicates Add-in (OECD, 2009c).

The third research question asks which students are advantaged or disadvantaged by using a YSD compared to an unbalanced design. For the sake of an efficient presentation of the results, the design-specific effects of several student variables on reading literacy were jointly analyzed with a multiple regression analysis. The weighted likelihood estimator (WLE; Warm, 1989) for performance in reading was used as the criterion variable. WLEs were used here to include only students in the analysis who answered reading items, which would not be the case for PVs. As predictors, several student variables known to have a strong relationship with reading performance were included in the analysis. The following effects were entered: booklet design (0 = unbalanced, 1 = original), immigration status (0 = native, 1 = immigrant, first and second generation; IMMIG), highest occupational status of parents (HISEI), gender (0 = female, 1 = male; ST04Q01), effort taking the test relative to maximum effort (CLCUSE3a), and all first level interactions of booklet design with the listed predictor variables. The codes given in the parentheses are the IDs of the variables as described in the German PISA 2006 scale documentation (Frey, Taskinen, et al., 2009). Further details of the calculation of the variables and the full item text can also be found in the scale documentation. School track (0 = other, 1 = Gymnasium/secondary school), and German federal state (dummy coded, reference category: Saxony-Anhalt) were included as control variables.

However, the multiple regression analysis could not be applied directly to the data sets for the conditions original and unbalanced. This is because the responses largely overlap and cannot be regarded as being independent from each other. To avoid this problem, a split-half approach was applied by drawing two distinct stratified random samples. The resulting subsamples are equivalent with regard to the variables gender and booklet, nested in school track, which in turn is nested in German federal states. The subsamples do not overlap and can thus be regarded as independent. By keeping the distribution of gender and booklet in school tracks in the German federal states equivalent between subsamples, no bias is caused by these variables in the multiple regression analysis. In order to receive unbiased standard errors, the student weights were calculated anew for both subsamples. This relatively unusual procedure allows examining research question three in a very efficient way and provides unequivocal results.

Results

As a prerequisite for comparing the results between the three booklet conditions, it has to be shown that all 3 (booklet design) × 16 (federal states) = 48 individual IRT based scalings resulted in a reasonable model fit and produced scales with sufficient reliability. The

calculated model fit and reliability indices are presented in the first part of the results section. Then, the observed effects of the variation in booklet design on key PISA results are shown. Specifically, the mean and gender differences in reading performance are reported for the 16 German federal states. Finally, the results of the multiple regression analysis are presented in order to analyze which students are likely to benefit from the YSD compared to an unbalanced design.

Model fit and reliability

The model fit of the three booklet design conditions is shown in Table 3. Since the deviance (-2 *Log-likelihood) depends upon the sample size, the variations in deviance between the original design and the other two designs as well as between the federal states should not be interpreted directly. Under the condition of a sufficiently large sample size with respect to the complexity of the scaling model in use, the ratio of the deviance and the sample size N can be used for comparisons. This ratio proved to be relatively stable between the German federal states, indicating a comparable global model fit. Thus, the results in the following sections are not influenced by a systematic lack of model fit.

Comparing the two booklet design conditions with similar sample sizes, the unbalanced design achieved a marginally better average model fit (Dev./N = 116) than the design with the random deletion of 2/13 of the responses (Dev./N = 118). Thus, it can be noted that an unbalanced booklet design cannot necessarily be identified by a lack of global model fit. This conclusion is supported by the reliabilities shown in Table 4. No substantial differences can be observed between the designs.

Mean performance

The unbalanced booklet design was specified in order to cause a maximum effect on the average student performance in reading. Since performance typically tends to decrease towards the end of the test, by deleting the responses of the two booklets with reading clusters in the last position, the student performance in reading was expected to be higher in the unbalanced condition compared to the other two conditions. The mathematics and science clusters are nearly equally distributed across the three remaining positions one, two, and three in the deleted booklets. Hence, no differences were expected between the booklet designs for mathematics and science. This assumption is confirmed by the results obtained. In no German federal state were significant differences in mathematics and science observed between the booklet designs. Most of the differences amount to less than one point on the PISA scale. Therefore, for the rest of the paper, only the results for reading are presented.

As expected, the average student performance in reading proved to be higher when the unbalanced design was used compared to the original booklet design and the design with the random deletion of responses (Table 5). In all federal states, the mean performance in

Table 3:
Model Fit by Booklet Design and Federal State

	Original			Unbalanced			Random Deletion		
Federal State	**Dev.**	***N***	**Dev./*N***	**Dev.**	***N***	**Dev./*N***	**Dev.**	***N***	**Dev./*N***
Saarland	219,157	1,834	119	182,898	1,557	117	185,004	1,546	120
Rhineland-Palatinate	277,423	2,343	118	229,477	1,980	116	234,651	1,986	118
North Rhine-Westphalia	297,668	2,474	120	245,678	2,082	118	251,513	2,097	120
Lower Saxony	205,450	1,732	119	169,786	1,462	116	173,493	1,465	118
Bremen	210,125	1,798	117	174,745	1,524	115	179,656	1,532	117
Schleswig-Holstein	337,281	2,866	118	281,123	2,435	115	286,527	2,440	117
Hamburg	400,743	3,369	119	333,151	2,854	117	337,435	2,832	119
Mecklenb.-Western Pomerania	218,138	1,828	119	180,894	1,544	117	185,526	1,556	119
Brandenburg	219,666	1,870	117	181,977	1,576	115	185,106	1,580	117
Berlin	323,888	2,783	116	268,678	2,355	114	275,382	2,362	117
Saxony	216,256	1,843	117	178,920	1,556	115	181,169	1,549	117
Bavaria	355,042	2,980	119	293,802	2,516	117	297,519	2,503	119
Baden-Württemberg	207,927	1,758	118	170,286	1,473	116	176,001	1,488	118
Hessia	449,261	3,793	118	373,076	3,209	116	382,996	3,227	119
Thuringia	218,243	1,874	116	180,711	1,585	114	187,573	1,617	116
Saxony-Anhalt	218,411	1,868	117	182,428	1,591	115	186,113	1,582	118
Mean	273,417	2,313	118	226,727	1,956	116	231,604	1,960	118

Note. Dev. = deviance (-2LogL), Dev./*N* = ratio of deviance and sample size. Students with special educational needs not included.

Table 4:
Reliability by Booklet Design and Federal State

	Original			Unbalanced			Random Deletion		
Federal State	**READ**	**MATH**	**SCIE**	**READ**	**MATH**	**SCIE**	**READ**	**MATH**	**SCIE**
Saarland	.852	.851	.893	.846	.871	.920	.846	.823	.892
Rhineland-Palatinate	.847	.878	.930	.846	.868	.908	.808	.835	.918
North Rhine-Westphalia	.777	.823	.887	.785	.856	.920	.803	.856	.902
Lower Saxony	.791	.839	.896	.737	.819	.887	.771	.824	.907
Bremen	.873	.877	.923	.819	.865	.889	.822	.883	.926
Schleswig-Holstein	.782	.859	.895	.772	.836	.892	.814	.848	.890
Hamburg	.805	.839	.896	.805	.843	.903	.808	.857	.902
Mecklenb.-Western Pomerania	.785	.847	.907	.834	.851	.906	.779	.812	.855
Brandenburg	.797	.877	.898	.758	.874	.887	.806	.887	.893
Berlin	.808	.859	.903	.794	.876	.919	.803	.867	.915
Saxony	.820	.818	.884	.771	.796	.873	.791	.832	.884
Bavaria	.781	.842	.880	.796	.861	.912	.809	.874	.906
Baden-Württemberg	.793	.834	.900	.820	.815	.878	.805	.842	.888
Hessia	.820	.868	.914	.782	.846	.895	.804	.854	.887
Thuringia	.830	.826	.903	.784	.847	.889	.860	.847	.924
Saxony-Anhalt	.782	.834	.886	.778	.880	.910	.803	.863	.884
Mean	.811	.849	.901	.797	.852	.900	.809	.852	.900

Note. EAP/PV reliabilities derived from five-dimensional model (reading, mathematics, science, interest in science, support for science). READ = reading literacy, MATH = mathematical literacy, SCIE = scientific literacy. Mean reliabilities were calculated using Fisher transformation.

Table 5:
Mean Student Performance in Reading by Booklet Design and Federal State

Original			Unbalanced			Random Deletion		
Federal State	***M***	***SE***	**Federal State**	***M***	***SE***	**Federal State**	***M***	***SE***
Saxony	**512**	2.9	Saxony	**517**	5.6	Saxony	**512**	5.9
Bavaria	**511**	3.4	Bavaria	**515**	4.7	Bavaria	**512**	5.5
Thuringia	**500**	4.0	Baden-Württemberg	**508**	4.8	Thuringia	500	4.9
Baden-Württemberg	500	4.2	Thuringia	**507**	6.7	Rhineland-Palatinate	499	3.6
Rhineland-Palatinate	**499**	3.0	Rhineland-Palatinate	**506**	4.9	Baden-Württemberg	497	5.7
Saarland	497	2.9	*Germany*	**501**	3.8	Saarland	497	5.2
Germany	495	1.5	Saarland	500	4.8	*Germany*	495	3.4
Hessia	492	3.5	Hessia	497	5.1	*OECD Average*	492	.6
OECD Average	492	.6	Berlin	495	5.3	Hessia	491	4.6
North Rhine-Westphalia	490	4.1	North Rhine-Westphalia	494	5.5	North Rhine-Westphalia	491	4.6
Berlin	488	3.7	Brandenburg	494	6.3	Brandenburg	490	5.5
Lower Saxony	487	3.8	Lower Saxony	492	6.5	Berlin	488	5.5
Brandenburg	486	4.9	*OECD Average*	492	.6	Lower Saxony	486	6.7
Schleswig-Holstein	**485**	3.0	Lower Saxony	491	5.2	Schleswig-Holstein	484	3.7
Lower Saxony	**484**	3.5	Schleswig-Holstein	491	4.1	Lower Saxony	**482**	4.9
Mecklenb.-Western Pomerania	**480**	5.3	Hamburg	487	7.1	Mecklenb.-Western Pomerania	**475**	5.5
Hamburg	**476**	5.5	Mecklenb.-Western Pomerania	486	4.9	Bremen	**474**	4.4
Bremen	**474**	4.0	Bremen	**481**	4.8	Hamburg	**473**	7.6

Note. Means differing significantly from the OECD average are printed in bold.

reading is higher for the unbalanced design than for the other two designs. The mean for Germany is about six points higher for the unbalanced design than for the original design. This difference is of a relevant magnitude: As a rule of thumb, the average increase in reading performance achieved over the course of one school year is frequently assumed to be around 30 to 40 points on the PISA scale. Thus, the observed difference equals the increment in reading literacy of about two months of schooling. It can further be seen from Table 5 that while the mean reading performance differs only slightly between the original design and the design with the random deletion of responses, the standard errors of the mean scores are substantially higher for the latter. This was also expected because the standard error of the mean is a function of the number of responses, which is about 2/13 lower for the design with the random deletion of responses. As a result of the higher standard errors of the mean, less countries show significant differences in reading performance to the average reading performance in the OECD area for the random deletion design compared to the original design.

Gender differences

Gender differences in reading are generally relatively large in PISA (e.g., OECD, 2010; Klieme et al., 2010). A comparison of the gender differences in reading performance between the original and the unbalanced design is shown in Figure 1.

	Girls		Boys		Difference (girls-boys)
	Original	Unbalanced	Original	Unbalanced	
Federal State	*M*	*M*	*M*	*M*	
Saarland	509	512	485	489	
Berlin	502	507	474	483	
Baden-Württemberg	514	518	486	498	
Hamburg	491	506	462	469	
Hessia	508	512	477	483	
Bavaria	531	534	494	498	
Mecklenb.-Western Pomerania	499	508	462	467	
Schleswig-Holstein	504	509	466	474	
Saxony-Anhalt	509	513	467	474	
Saxony	533	538	490	495	
North Rhine-Westphalia	512	515	469	474	
Bremen	496	502	453	461	
Thuringia	521	530	477	483	
Brandenburg	510	512	463	477	
Rhineland-Palatinate	521	526	474	485	
Lower Saxony	508	516	459	465	

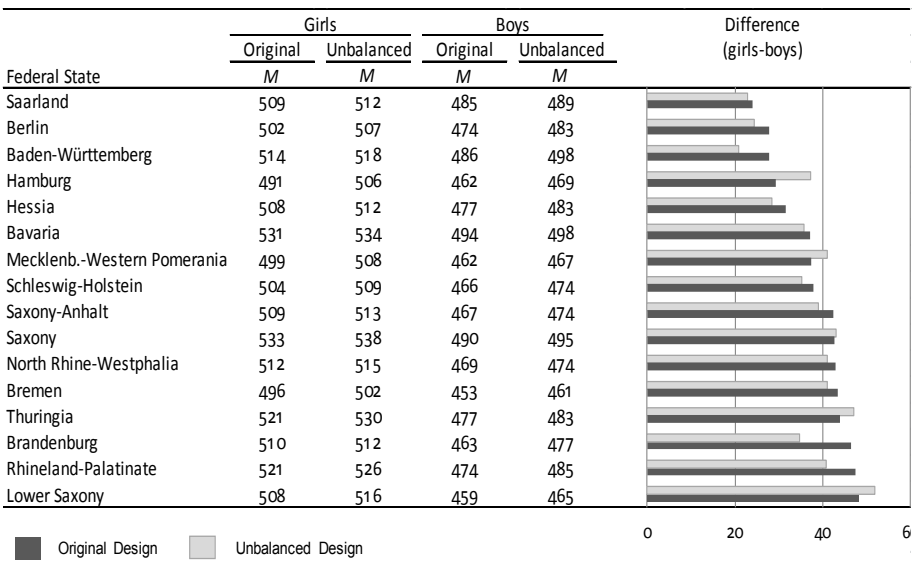

Figure 1:
Differences in Reading Performance between Boys and Girls by Booklet Design and Federal State.

Besides the effect that reading performance is generally higher for the unbalanced design than for the original design, it can be seen that the shifts caused by the change in booklet design differ between federal states. While in Brandenburg the gender difference declined from 47 points to 35 points (-12 points), it increased from 29 points to 37 points in Hamburg (+8 points). Thus, unbalanced booklet designs may not only cause main effects on average performance, but can also induce differentiated effects on subpopulations.

Interactions between student variables and booklet design

The results presented so far have shown effects of using an unbalanced instead of a balanced booklet design on the reading performance distribution in PISA-E 2006. To foster a more pronounced understanding of the phenomenon, the third research question asks whether students with special characteristics will benefit when a balanced booklet design is used instead of an unbalanced booklet design. A multiple regression analysis was carried out to provide an answer to this question using a set of student variables known to have a strong relationship with reading performance.

Table 6:
Multiple Regression Analysis Predicting Reading Performance by Booklet Design and Individual Student Characteristics

Variable	***B***	***SE***
Design	-6.67^{+}	3.48
Immigration Status	-38.51**	4.38
HISEI	15.01**	1.89
Gender	-30.88**	5.31
Effort	8.22**	1.10
Design x Immigration Status	-12.04**	3.99
Design x HISEI	2.12	2.76
Design x Gender	1.58	4.27
Design x Effort	5.11**	1.26
R^2	0.31**	0.01

Note. N = 14,193. HISEI = highest occupational status of parents. Design, immigration status and gender are dichotomous variables. Effort and HISEI included as z-standardized variables.

Federal state (dummy coded) and school track (other vs. Gymnasium) were used as control variables.

*$p < .05$, two-tailed. **$p < .01$, two-tailed. $^{+}p < .05$, one-tailed. $^{++}p < .01$, one-tailed.

As can be seen in Table 6, all of the main effects of the predictor variables on the criterion variable reading performance are significant ($p < .05$). The difference in reading performance between the original design and the unbalanced design is of comparable magnitude as reported in the previous section. Note that a one-tailed test was applied for the main effect of the predictor variable design on reading performance because lower reading performance for the original booklet design (coded as 1) compared to the unbalanced booklet design (coded as 0) was a-priori assumed. The estimate of -6.67 for the regression coefficient of the predictor variable design means that the reading performance in the balanced design condition was estimated to be 6.67 points lower than for the unbalanced design. Furthermore, and in line with official PISA results, male students with an immigration background and a low socio-economic background tend to have relatively low performance scores in reading. Additionally, the effort spent on working on the test has a significantly positive relationship with the performance in reading.

Two of the interaction effects are significant. The significant Design × Immigration Status interaction means that the average reading performance of students with an immigration background is about 12 points lower for the original booklet design than for the unbalanced booklet design. The significant Design × Effort interaction denotes that a shift of one standard deviation in the variable effort will go hand in hand with reading performance being five points higher for the balanced design than for the unbalanced design. Thus, students without an immigration background and students reporting high levels of effort will benefit from using the YSD design instead of the unbalanced design.

Discussion

The study addresses the question of which problems are avoided by using a YSD instead of an unbalanced booklet design in PISA. It is examined whether using booklet information in the scaling procedure removes the problems a YSD can control for or whether both – booklet correction within scaling and a YSD – are necessary in order to optimize the interpretability of PISA results.

First of all, the results underline the fact that an unbalanced booklet design cannot necessarily be detected by a lack of model fit or reliability. Model fit and reliability indices are not based on assumptions about the balancing of underlying design structures and are therefore not sensitive for detecting unbalanced designs. Correspondingly, at best, marginal differences in model fit and reliability between booklet designs are observed in the present study. In fact, a better model fit can sometimes be expected for unbalanced designs than for balanced designs if an unbalanced design systematically diminishes effects violating the assumptions of the IRT model used for scaling. This would be the case, for example, if position effects are expected to be especially large at the end of a test and an unbalanced design is built by deleting the last item clusters from all the booklets of a balanced design. If position effects are present, the data assessed with the resulting unbalanced design will be less prone to model violations due to position effects and thus more likely to achieve a better model fit compared to a model based on the data assessed

with the original balanced design. A tendency towards this effect pattern was observed in the present study.

Furthermore, the results of the present re-analysis clearly demonstrate that distortions of the balanced booklet structure can cause substantial and relevant shifts in the estimated performance distribution at the level of subpopulations when position effects are present, which is typically the case for LSAs. On average, the mean student performance in reading was about six points higher for the unbalanced design compared to the balanced design. Complicating things, strong evidence was found that the effects induced by altering the design structure differ between subpopulations. For example, the differences observed between the reading performance of boys and girls decreased in some German federal states when an unbalanced booklet design was used instead of the original booklet design, while it increased in others. Furthermore, the results suggest that students with an immigration background show a significantly lower performance in reading when the original booklet design is used, compared to the unbalanced design. With the data assessed in PISA, a sound explanation for these differentiated yet systematic effects cannot be given. In fact, in most cases, it would be impossible to explain such a variety of effects without carrying out analyses explicitly designed for their examination. Nevertheless, the observed effects show that using an unbalanced booklet design causes severe problems regarding the interpretability and thus the validity of results on the level of subpopulations. To control for these systematic effects, it is thus definitely a good idea to use a balanced booklet design in order to control for position effects. An additional approach worth considering is using a more complex IRT model for scaling. The observed differences between the balanced and the unbalanced design indicate that applying a booklet correction might not be sufficient to account for the position and carry-over effects present in PISA data.

Besides the effects mentioned, the perceived effort spent on answering the test items interacts with the type of booklet design. A difference in perceived effort of about one standard deviation is associated with an observed difference in reading performance of about five points when the original design was used compared to the unbalanced design. Even though this effect is relatively small, it raises some interesting general considerations regarding validity. Before going into detail, it has to be noted that the significant Effort × Design interaction does not necessarily mean that effort has a causal effect on performance in reading, since the effort spent was assessed after working on the test items in PISA 2006. In fact, it seems more likely that (perceived) test performance has an influence on low effort ratings. However, the observed lower performance for the unbalanced design compared to the balanced design (which is assumed to be due to fatigue effects or a decline in test taking motivation) in combination with the Effort × Design interaction indicates that under the balanced design, the measured reading performance falls below the maximum reading performance of the students. Thus, it is important to consider whether PISA fully achieves its aim to measure *what students know and can do in specified content areas*. In fact, the results of the present study suggest that PISA does not measure maximum performance but rather *what students know and can do in specified content areas within a testing time of two times 60 minutes interrupted by a break of 15 minutes*. Thus, if measuring the actual maximum performance is the aim, shortening

the testing time would probably foster the validity of the results. However, if measurement precision should be kept at the present level, shortening the individual testing time would mean testing more students or shifting to more efficient administration procedures like multi-stage testing or computerized adaptive testing. Using the PISA item pool, Frey and Seitz (2011) and Frey, Seitz and Kröhne (2013) have recently shown that the average number of items that need to be presented to the students can be substantially reduced by using multidimensional adaptive testing, even if the restrictions connected with PISA are taken into account. Therefore, computerized adaptive testing seems to be an option even though it would mean moving from paper & pencil administration to computer-based assessment.

Summing up, there are five major lessons that can be learned from the present study.

1. Unbalanced booklet designs cannot necessarily be detected by a lack of model fit or a lack of reliability.
2. Using unbalanced booklet designs can have a severe impact on population estimates of student achievement in large-scale assessments.
3. The effects of using an unbalanced booklet design can differ between subpopulations.
4. A change in booklet design is accompanied with systematical advantages or disadvantages for some students.
5. The testing time of PISA may be too long for a valid measurement of maximum performance.

Nevertheless, the study has some limitations and thus leaves room for further research. First of all, the analyses are based on the data from the German PISA-E 2006 assessment and are therefore of limited generalizability. Different results might be observed for other countries or other assessments. Consequently, a replication of the present study including all countries participating in PISA would be interesting. Furthermore, and as discussed above, a shorter testing time may reduce the differences between the balanced and the unbalanced booklet design. The examination of data stemming from other LSAs with shorter testing periods would therefore provide further insights into psychological test-taking processes. Lastly, since the focus of the present study lies on the impact of using a YSD on population estimates in PISA, the results do not provide a differentiated view of position effects even though they are likely to be the cause of the differences observed between the unbalanced and the original booklet design. More in-depth analyses directly addressing position effects in PISA can be found in the PISA technical reports and the paper of Hartig and Bucholz (2012) in this issue.

In conclusion, the results underline the importance of using a balanced booklet design for large-scale assessment of student achievement. A balanced booklet design helps to avoid or minimize unwanted bias in the population estimates of interest, even though the underlying psychological processes are not yet fully understood. We hope that this paper stimulates more research in the largely neglected field of booklet designs.

Funding note

The preparation of this article was supported in part by grant 01DB1104 (MaK-adapt) from the German Federal Ministry of Education and Research (BMBF) within the initiative "Innovative skills and competence assessment to support vocational education and training" (ASCOT)".

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

Adams, R. J., Wilson M., Wu, M. (1997). Multilevel item response models: an approach to errors in variable regression. *Journal of Educational and Behavioral Statistics*, *22*, 47–76.

Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D.A. (1962). *Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959–1961*. Hamburg: UNESCO Institute for Education.

Frey, A., Carstensen, C. H., Walter, O., Rönnebeck, S., & Gomolka, J. (2008). Methodische Grundlagen des Ländervergleichs [Methods of the comparison between the German federal states]. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, & R. Pekrun (Eds.), *PISA 2006 in Deutschland: Die Kompetenzen der Jugendlichen im dritten Ländervergleich* (pp. 375-397). Münster: Waxmann.

Frey, A., Hartig, J., & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: theory and practice. *Educational Measurement: Issues and Practice*, *28*, 39-53.

Frey, A., & Seitz, N. N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in PISA. *Educational and Psychological Measurement*, *71*, 503-522.

Frey, A., Seitz, N. N., & Kröhne, U. (2013). Reporting differentiated literacy results in PISA by using multidimensional adaptive testing. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research on PISA*. Dodrecht: Springer.

Frey, A., Taskinen, P., Schütte, K., Prenzel, M., Artelt, C., Baumert, J., … Pekrun, R. (Eds.). (2009). *PISA 2006 Skalenhandbuch. Dokumentation der Erhebungsinstrumente* [PISA 2006 scale documentation]. Münster: Waxmann.

Giesbrecht, F. G., & Gumpertz, M. L. (2004). *Planning, construction, and statistical analysis of comparative experiments*. Hoboken, NJ: Wiley.

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.

Hartig. J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, *54,* 418-431.

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: An International Journal on Theory and Practice, 17*, 497-509.

Jones, L. V., & Olkin, I. (Eds.). (2004). *The nation's report card: evolution and perspectives*. Bloomington: Phi Delta Kappa Educational Foundation.

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., … Stanat, P. (Eds.). (2010). *PISA 2009. Bilanz nach einem Jahrzehnt* [PISA 2009. Balance after one decade]. Münster: Waxmann.

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.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Chestnut Hill, MA: Boston College.

Mullis, I. V .S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: Boston College.

OECD (2002). *PISA 2000 technical report.* Paris: OECD Publishing.

OECD (2005). *PISA 2003 technical report*. Paris: OECD Publishing.

OECD (2009a). *PISA 2009 assessment framework – key competencies in reading, mathematics and science*. Paris: OECD Publishing.

OECD (2009b). *PISA 2006 technical report*. Paris: OECD Publishing.

OECD (2009c). *PISA data analysis manual. SPSS®* (second edition). Paris: OECD Publishing.

OECD (2010). *PISA 2009 results: what students know and can do: student performance in reading, mathematics and science (volume I)*. Paris: OECD Publishing.

OECD (2012). *PISA 2009 technical report*. Paris: OECD Publishing.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R. (Eds.). (2008). *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich* [PISA 2006 in Germany: Third comparison of students‘ competences between the German federal states]. Münster: Waxmann.

Stanat, P., Pant, A., Böhme, K., & Richter, D. (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2012* [Student competencies in German and mathematics at the end of grade four]. Münster: Waxmann.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Way, W. D., Carey, P., & Golub-Smith, M. (1992). *An exploratory study of characteristics related to IRT item parameter invariance with the Test of English as a Foreign Language (TOEFL Tech. Rep. No. 6.)*. Princeton, NJ: Educational Testing Service.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Melbourne, Victoria, Australia: ACER Press.

Yousfi, S., & Böhme, H. (2012). Principles and procedures of considering context effects in the development of calibrated item pools: Conceptual analysis and empirical illustration. *Psychological Test and Assessment Modeling*, *54*, 366-396.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, *10*, 10–16.