Educational Measurement: Issues and Practice
Fall 2009, Vol. 28, No. 3, pp. 39–53

An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice

Andreas Frey, *Leibniz Institute for Science Education (IPN),* Johannes Hartig, *University of Erfurt,* and André A. Rupp, *University of Maryland*

In most large-scale assessments of student achievement, several broad content domains are tested. Because more items are needed to cover the content domains than can be presented in the limited testing time to each individual student, multiple test forms or booklets are utilized to distribute the items to the students. The construction of an appropriate booklet design is a complex and challenging endeavor that has far-reaching implications for data calibration and score reporting. This module describes the construction of booklet designs as the task of allocating items to booklets under context-specific constraints. Several types of experimental designs are presented that can be used as booklet designs. The theoretical properties and construction principles for each type of design are discussed and illustrated with examples. Finally, the evaluation of booklet designs is described and future directions for researching, teaching, and reporting on booklet designs for large-scale assessments of student achievement are identified.

Keywords: booklet design, experimental design, item response theory, large-scale assessments, measurement

Andreas Frey is a Senior Researcher at the Leibniz Institute for Science Education (IPN) at the University of Kiel, Olshausenstr. 62, D-24098 Kiel, Germany (frey@ipn.uni-kiel.de). His primary research interests include multidimensional adaptive testing, booklet designs, and methodology of large-scale assessments. Johannes Hartig is a Full Professor, Faculty of Education, Department of Educational Research Methodology, University of Erfurt, POB 900221, 99105 D-Erfurt, Germamy (johannes.hartig@uni-erfurt.de). His primary research interests include item difficulty modeling and multidimensional item response theory. André A. Rupp is an Assistant Professor in the Department of Measurement, Statistics, and Evaluation (EDMS) at the University of Maryland, 1230 Benjamin Building, College Park, MD 20742 (ruppandr@umd.edu). His research interests include the theory and practice surrounding diagnostic classification models vis-à-vis alternative explanatory item response theory models and the development of modern measurement methods for games- and simulation-based environments.

Series Information

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education. Information regarding the development of new ITEMS modules should be addressed to: Dr. Mark Gierl, Canada Research Chair in Educational Measurement and Director, Centre for Research in Applied Measurement and Evaluation, Department of Educational Psychology, 6-110 Education North, University of Alberta, Edmonton, Alberta, Canada T6G 2G5.

Studies that use *large-scale assessments of student achievement* are a ubiquitous component of systems monitoring in education. Internationally, some of the most well-known studies include the *Programme for International Student Assessment (PISA)*, the *Trends in International Mathematics and Science Study (TIMSS)*, and the *Progress in International Reading Literacy Study (PIRLS)*. Nationally, the *National Assessment of Educational Progress (NAEP)* in the United States and the *School Achievement Indicators Program (SAIP)* as well as its successor, the *Pan-Canadian Assessment Program (PCAP)* in Canada, supplement such international studies. Moreover, many provinces or states within a given country conduct independent large-scale assessments of student achievement that are administered to a representative sample or the entire population of students at regular intervals.

The objective of these assessments is to obtain reliable information about student achievement in one or more content domains of interest. Typically, several hundred questions are used to test the respective content domain(s). Administering all questions to each student participating in such a study would be too time consuming. Therefore, students are given different test forms—*booklets*—each containing a number of questions that can be sensibly answered by a student in the available testing time. The way the questions are assigned to the booklets is specified by a so-called *booklet design*.

Because poorly chosen booklet designs can lead to substantial bias in both the estimation of item parameters and ability distributions, the choice of an adequate booklet design is highly relevant to ensure reliable test scores and their valid interpretation. The key question that arises is, thus, how a booklet design can be constructed that can prescribe how to administer the items in order to obtain unbiased and efficient parameter estimates.

Given the widespread use of large-scale assessments of student achievement internationally and nationally, one would expect a rich literature to exist that provides systematic and coherent reviews of the key theories that underlie the booklet designs of these assessments. One would also expect that the procedural expertise that has been gathered by assessment designers in the testing committees for these large-scale assessments has already been captured in a wide variety of practical guidelines that facilitate this complex endeavor for other colleagues. Surprisingly, this is not the case.

Put simply, there is no "theory of booklet design" that is documented in a satisfactory manner so that it can guide specialists in the process of constructing a booklet design that is optimal for a particular large-scale assessment. The type of specialist that would need this kind of information is most likely someone with a solid training in measurement but without specific expertise in the construction of booklet designs. The few existing papers explicitly addressing booklet designs merely discuss the topic superficially (e.g., Beaton & Zwick, 1992; Childs & Jaciw, 2003). For large-scale assessments of student achievement, *technical reports* are published that describe the key steps taken in their development (e.g., Allen, Donoghue, & Schoeps, 2001; Martin, Mullis, & Kennedy, 2007; OECD, 2009; Olson, Martin, & Mullis, 2008; see also Rupp, Vock, Harsch, & Köller, 2008). The technical reports, however, tend to focus on the final booklet design that was used for the assessment. They typically do not describe the myriad choices that needed to be made in the generation of such a design, the compromises that these choices entailed, the alternative designs that could also have been utilized, and the reasons for why they were not selected. In short, they do not provide didactic support for the replication of such steps in novel contexts as the reports are of a referential, not educational, nature and are oriented toward the product and not the process of its genesis.

Similarly, textbooks on educational measurement or state-of-the-art reference volumes such as the *Handbook of Test Development* (Downing & Haladyna, 2006), *Educational Measurement*, Fourth Edition (Brennan, 2007), or the *Handbook of Statistics, Volume 26: Psychometrics* (Rao & Sinharay, 2007) do not cover booklet designs in detail either. However, the implications of booklet designs are addressed when issues such as equating and linking, model estimation, and sampling are discussed. This shows a large disconnect between the actual practice of booklet design and the theoretical importance that it has for essentially all subsequent scaling and reporting stages of a large-scale assessment of student achievement.

From a theoretical perspective, two scientific disciplines deal with principles that are very useful for booklet designs. These disciplines are *combinatorics* as part of mathematics (e.g., Tucker, 2006) and *experimental design* as part of statistics (e.g., Giesbrecht & Gumpertz, 2004). However, applying the principles of combinatorics and experimental design to the context of booklet design for large-scale assessments of student achievement is no trivial task as they have historically evolved within quite different contexts with quite different objectives. Combinatorics is concerned with the theoretical issues of selection, arrangement, and combination of objects chosen from a finite set. Mostly, even work that is considered applied in this area remains rather abstract from the perspective of practical booklet designs. Within the field of experimental design, methodological research is predominantly oriented toward data structures arising from traditional laboratory or field studies with observed variables. Even though one can view large-scale assessments as educational surveys, it is not straightforward to extend the statistical models for experimental designs to accommodate the specific survey structure and the associated requirements of latent-variable models (for a discussion see, e.g., Mislevy & Rupp, 2009, unpublished data). Thus, although the number of designs and design features within the experimental design literature is extensive, it is rather challenging to select a design that is directly utilizable as a booklet design, given the restrictions a particular large-scale assessment entails.

In sum, measurement specialists in expert committees who are responsible for constructing booklet designs are forced to learn about the process of booklet design from (a) the technical reports of previous large-scale assessments, (b) general descriptions of methods in combinatorics and experimental design, or (c) explanations from colleagues who have had experience with these designs in the past. The present module was conceived to overcome this rather dissatisfying lack of support for solving this complex task. It introduces the construction of booklet designs as the task of finding a solution for allocating a large number of questions to a smaller number of booklets under context-specific constraints. The module is written in an accessible language so that it can serve as a *didactic consciousness-raising device* that illustrates key principles, potential decisions and their resulting consequences, and provides measurement specialists with a toolbox and a structured way of thinking about the problem.

Because the needs of individual assessment contexts can induce unique constraints into the process of booklet design, this module obviously cannot provide answers for all potential design questions. Nevertheless, besides providing designs for typical large-scale assessment situations, it will empower the reader to ask more targeted questions and to devise solutions that are based on sound measurement principles grounded in broadly accessible disciplinary expertise rather than the artistic ability of a chosen few.

What will not be talked about in this module, however, is the process of automatically generating unique *optimal designs* with the help of search algorithms that do not fit a particular typical design structure (see, e.g., van der Linden, Veldkamp, & Carlson, 2004; Verschoor, 2007). Nor does it involve an in-depth discussion of the choice of *measurement models* available to the practitioner. Thus, it will not suggest a particular measurement model as superior to others, as the choice of a measurement model is context-specific and depends on resource constraints, available expertise, and measurement traditions as well. Furthermore, excellent overviews of measurement models for large-scale assessments already exist (e.g., de Ayala, 2009; de Boeck & Wilson, 2004; Embretson & Reise, 2000; McDonald, 1999; Muthén, 2002; Pellegrino, Chudowsky, &

Glaser, 2001; Rupp & Templin, 2008; Rupp, Templin, & Henson, in press; Thissen & Wainer, 2001; von Davier, Sinharay, Oranje, & Beaton, 2006).

The module is divided into four main sections. In the first section, the major *statistical objectives* of booklet designs for large-scale assessments of student achievement are summarized as they lay the foundation for understanding why booklet design is so critical to data analysis and reporting. In the second section, critical *constraints* that influence the structure of a booklet design within a particular application context are reviewed. In the third section, several *types of designs* that are available for the process of booklet design are introduced. In the fourth section, indices for the *evaluation* of designs are presented. The module closes with a brief summary, an outlook for possible research and teaching directions on this topic, and a call to other measurement experts to pursue them.

Statistical Objectives of Large-Scale Assessments of Student Achievement

In order to be able to talk clearly and coherently about booklet designs, a few recurring terms need to be defined at the outset. These include item, testlet, item pool, item cluster, and booklet. Specifically, we will refer to a single question in an assessment as an *item* and a set of questions that is connected via a common stimulus such as a text in a reading comprehension assessment or a graph in a mathematics assessment as a *testlet* in alignment with the mainstream measurement literature (e.g., Wainer, Bradlow, & Wang, 2007). To accelerate the readability of the module, in the following we mostly speak of items and only differentiate between items and testlets when it is necessary. The overall collection of items from which a selection for the assembly of the different booklets has to be made will be referred to as the *item pool*. We will use the term *item cluster* or just *cluster* to refer to subsets of the item pool that are presented together. The item clusters are mostly used as building blocks for a particular test form that a student responds to; such a test form is referred to as a *booklet*. Booklets typically contain items from one or more content domains (e.g., reading, mathematics, and science).

In order to understand the decisions that developers of large-scale assessments of student achievement make when constructing a booklet design, it is useful to differentiate between the following two statistical objectives:

1. Calibration of the item pool (i.e., estimation of item parameters).
2. Proficiency scaling of students (i.e., estimation of person parameters).

Under both objectives the main concern is to obtain unbiased and efficient estimates of the parameters of interest so that the quality of the assessment and the proficiency of student groups can be gauged reliably.

Consequently, response data need to be analyzed with a measurement model that provides such estimates within the context of a relatively complex design structure. The most commonly used measurement models for large-scale assessments of student achievement are unidimensional or multidimensional models from item response theory (IRT) (see, e.g., de Ayala, 2009; Embretson & Reise, 2000; van der Linden & Hambleton, 1997; Yen & Fitzpatrick, 2006; or the respective chapters in the technical reports of NAEP, PISA, TIMSS, and PIRLS mentioned above). IRT models locate item and person parameters on the same latent scale, which allows for an interpretation of person parameters in terms of the probability with which they can solve specific items. The mathematical separation of item and person parameters in IRT models holds numerous theoretical and practical advantages that are important for large-scale assessments of student achievement. For example, they allow for the equating of scores across multiple test forms and time points, the direct estimation of potential biases across latent or manifest student groups, and the automatic selection of items from item pools using computer-adaptive schemes.

A *calibration of the item pool* is typically of primary concern in early phases of the data analysis when the psychometric properties of the instruments need to be established. Item parameters that are of major concern include difficulty, discrimination, or pseudo-guessing parameters. In order to investigate the functioning of the items accurately, parameters should thus be estimated as precisely as possible. This requires, in turn, that a large number of students from the entire ability range for each subpopulation—but especially the entire range of ability that the item pool targets most—respond to each item because that will increase the efficiency of the estimates.

In essence, assessment developers use item parameters to identify those items that are not functioning as intended in order to remove them or recommend them for revision. This could mean, for example, that they are either too easy or too difficult, do not discriminate enough in certain ranges of the proficiency continuum, have guessing probabilities that are too high, contain some distractors that do not function as intended, or induce undue disadvantages for certain subpopulations of students created by variables such as sex, ethnicity, school type, or socioeconomic status.

A *proficiency scaling of students* is the main objective of large-scale assessments of student achievement and typically follows the calibration of the item pool. The major aim of large-scale assessments of student achievement lies on group-level/aggregate reporting, not on reporting at an individual level. Thus, booklet designs have to assure that statistics on the level of countries, districts, schools, or classrooms, rather than at the level of individual students, can be validly interpreted. For example, the measurement precision of individual proficiency estimates can be merely moderate in large-scale assessments of student achievement as long as the measurement precision of group-level statistics—predominantly, means and mean differences—is acceptably high.

To draw valid inferences about groups of students' proficiencies from their person parameter estimates, these estimates need to be similarly unbiased and as precise as possible over the entire range of ability that the assessment is designed for. This may be jeopardized by using an inappropriate booklet design. A simple problematic case would be if booklets with varying numbers of items are systematically used in specific subsamples that differ on secondary confounding variables that have a significant impact on achievement. In an international study of student achievement, this might be the case if shorter booklets are used in some countries and longer booklets are used in others, which could be motivated by a desire to make the assessment fit into the typical length of school sessions across different countries (e.g., 45 minutes vs. 60 minutes). Varying booklet length will not only result in differences in the average efficiency of the person parameters (because the number of data points have

a direct impact on the efficiency of the parameters) between the countries, but also in a possible estimation bias due to additional factors such as fatigue or reduction in test-taking motivation for the countries in which the longer booklets were used. To control for these problems, booklet designs in which booklets have an equal number of items should be used across all potentially relevant subgroups that are known a priori.

Design Constraints

In this section, key constraints for booklet designs are described, which induce challenges for realizing the statistical objectives of large-scale assessments of student achievement that were described in the previous section. The process of constructing a booklet design for a large-scale assessment of student achievement is essentially the task of systematically assigning items, testlets, or clusters of items to different booklets under a variety of practical constraints. Some of the most common factors that constrain booklet designs are:

1. The number of items in the item pool.
2. The number of content domains that need to be tested by the assessment.
3. The administration format of the assessment.
4. The testing time allotted for the administration of the items.
5. The possibility of position and carryover effects.
6. The planned linking with other assessments.
7. The need to keep items secret.

This list is, by no means, exhaustive and different assessment scenarios may induce additional constraints that do not seem to fit neatly into one of these categories. Nevertheless, the list covers the principal factors that strongly influence most booklet designs in practice.

Size of the Item Pool

One of the most critical factors influencing the range of applicable booklet designs is the *number of items in the item pool*. Of course, if the item pool was relatively small, the issue of constructing a booklet design would not arise because only one booklet or a small number of booklets could be administered to students. In large-scale assessments of student achievement the number of items in a pool is generally very large, however, and can consist of several thousand items, so that it is necessary to present students only subsets of the complete item pool.

If items are naturally grouped into testlets, such as in the case when multiple items are attached to a reading comprehension passage as in PISA or in PIRLS or when multiple items are attached to a statistical graphic as in TIMSS, it has to be kept in mind that the smallest unit that can be assigned to booklets is a testlet and not an item. Generally speaking, when the number of items or testlets becomes large, the booklet design becomes easier if items or testlets are grouped into clusters and the clusters are subsequently assigned to booklets.

Multiple Content Domains

The *number of content domains that need to be tested by the assessment* also influences the choice of the booklet design. A booklet design gets significantly more complicated if multiple content domains are covered, which is the case in studies such as PISA, TIMSS, or NAEP. If the assessment intends to measure several content domains with different degrees of breadth but a comparable precision across study cycles, a suitable booklet design for any given cycle needs to allocate the number of items per content domain proportionally in alignment with these objectives.

Administration Format

The *administration format of the assessment* also plays an important role. Theoretically, a booklet design may actually not be necessary at all for a large-scale assessment of student achievement because a computer-adaptive algorithm (e.g., Segall, 2005 or Wainer, 2000 for the one-dimensional case, or Frey & Seitz, in press for the multidimensional case) could be implemented that essentially creates a different booklet in real time for each student. However, a paper-and-pencil administration is still the most common assessment format for large-scale studies and all assessments listed at the beginning of the module predominantly use this format. This creates restrictions for the booklet design that are of a logistical nature. Specifically, a large number of booklets results in a high workload in four stages of the assessment.

First, it is costly to format and print many different booklets. Second, many different booklets result in a more complex test administration process. For example, it has to be ensured that the booklet each individual student responds to is the correct booklet. Third, the identification numbers of different booklets need to be properly tracked while collecting and scoring the response data, which can be especially challenging for longitudinal studies. Fourth, the process of scoring responses, recording scores, and analyzing the scores using a common data matrix is more complex and more resource intensive if many different booklets are used. Therefore, if tests are administered in a paper-and-pencil format, a lower number of booklets will generally be desirable due to economical reasons and in order to reduce sources of error in the data structure at different stages of the study that might lead to biased parameter estimates.

Testing Time

The *testing time allotted for the administration of the items* also has an impact on the range of possible booklet designs. The testing time primarily determines the number of items that can be presented within each booklet. For example, if an assessment is scheduled with 100 minutes pure testing time excluding instructions and breaks, and items in the item pool take about 4 minutes to complete, on average, then a design can be chosen that specifies booklets with a length of 25 items each. In PISA 2006, for example, each booklet contained four clusters of items with each cluster being allocated 30 minutes for a total of 120 minutes total testing time.

Typically, available items differ in the time that is required to answer them. In this case, clusters of items that require approximately the same testing time can be built. These clusters can then be used to easily assemble booklets that need approximately the same time to be answered. If the time that is allocated for particular clusters is insufficient for certain student groups, however, missing data on items at the end of the cluster would result from this design flaw, which would lead to biased parameter estimates.

Position Effects

One major cause of biased item parameter estimates are *position effects* of clusters in different booklets as well as the position effect of items within a cluster. Correct answers may be given more frequently if an item is presented in a cluster at the beginning of a booklet rather than at the end of a booklet due to growing fatigue, reduced test-taking motivation, or just a lack of time on the part of the student for longer booklets. For example, in PISA 2003, the item difficulty estimates were around .5 logits lower if the cluster containing the respective item was presented in the first position of a booklet compared to the last position of a booklet (OECD, 2005); similar position effects are reported for PISA 2006 (Le, in press). For an item of average difficulty, this means that the percentage of correct answers is about 10% higher if it appears at the beginning of a booklet compared to a presentation at the end of a booklet. Thus, in the first case, the same item appears to be easier than in the latter case although the difference in difficulty is only due to the variation of its cluster position and not due to the cognitive demands of the item.

If unaccounted for, such variations in item parameter estimates can lead to severe validity problems if the item difficulties are used to give meaning to sections of a scale, which can transform the problem from a measurement problem to a reporting problem. For example, proficiency levels that are determined by standard-setting procedures (e.g., Zieky & Perie, 2006) are often characterized based on the cognitive demands of items that are estimated to be at that level. If the relative position of the items is inaccurate due to biased item parameter estimates arising from position effects, inaccurate characterizations of the scale may arise.

To avoid position effects, the best solution would be to use short booklets. Unfortunately, this is not always possible. In assessment situations where relatively long testing sessions cannot be avoided, position effects can statistically be controlled for with a booklet design that presents all items in all possible positions with equal frequency. Thereby, the position effects are averaged out over the positions.[1] However, if the average of the effects of the different positions on item parameter estimates does not equal zero, strictly speaking, criterion-referenced inferences can only be drawn with respect to the item difficulties (or response probabilities) within the given test length of the assessment, or after an adequate statistical modeling of the position effects.

Carryover Effects

Another factor that may affect the accuracy of item parameter estimates is the context in which items are presented. If, for instance, a particular algebra cluster is presented after the student has already worked on a number of similar algebra clusters, he or she will probably find it easier than if the same cluster is presented after working on a number of geometry clusters or even distinct clusters from a different content domain such as reading. We will refer to the impact that the cluster or item context has on the response probabilities of subsequent items as a *carryover effect*, which is a specific type of position effect. When the concern is about the effect of a particular cluster on the following one, this is also referred to as a *first-order carryover effect,* whereas effects across more positions are referred to as *higher-order carryover effects*.

Carryover effects represent a serious problem for IRT-based assessments because they violate the assumption of local independence. Thus, at the stages of test construction and field testing, residual response dependencies due to carryover effects should be rationally identified, empirically quantified, and eliminated as far as possible. If concerns remain that carryover effects may occur for operational testing stages, the impact on parameter estimates should be statistically controlled for with a booklet design that balances combinations of clusters with regard to the characteristic that may influence responses to subsequent clusters. Like mentioned above for position effects, criterion-referenced inference should then, strictly speaking, only be drawn with respect to the item difficulties (or response probabilities) within the specific assessment at hand, or after an adequate psychometric modeling of the carryover effects.

Linking

Many large-scale assessments aim to report trends by comparing the results in the measured content domains between assessments carried out in different years. To justify the comparability of the test results, a *linking between assessments* (e.g., Kolen & Brennan, 2004) needs to be established. The linking is often done by using a set of so-called anchor items that is common between one or more assessments. If position effects and/or carryover effects have to be expected, a booklet design should be used that keeps the position and/or the sequence of the items constant over assessments. Establishing vertical scales with sufficiently strong linkages without compromising construct comparability over time is a challenging endeavor, however, whose nuances are beyond the scope of this module (see, e.g., Briggs, 2009).

Item Security

Many large-scale assessments of student achievement require that most or all operational items in the pool be kept secret. *Item security* is especially critical for studies whose results are of high political or economical relevance. In these cases, the test results can be jeopardized if items are memorized by participants and reported to persons taking the test later. This behavior can lead to an overestimation of students' proficiency, especially if a small number of items is presented in a single booklet. To reduce the feasibility of memorizing items, many items can be distributed to a relatively large number of booklets where they are presented in different clusters and different orders.

Booklet Design with the Aid of Experimental Designs

In this section, the terminology of experimental design will be adapted to the specific context of booklet design, and a variety of experimental designs that are suitable for booklet design will be reviewed. Common experimental designs like completely randomized designs are not discussed here as they are mostly not suitable as booklet designs. The following list of designs is by no means exhaustive as there are many basic designs and design modifications that may be suitable in a particular situation. In this sense, constructing a booklet design will be described based on key principles that go into that process rather than the mere selection of

an appropriate existing design. The latter would be an undue oversimplification of the complexities of the real-life scenarios that designers in large-scale assessments of student achievement are faced with.

The experimental designs that will be reviewed include *complete permutation designs (CPDs), balanced incomplete block designs (BIBDs)*—with a focus on the special case of *Youden squares designs (YSDs)*—and *repeated treatment designs (RTDs)*. For each design that is described, an example design table is provided, relevant practical uses of the design are outlined, the principles of construction are explained, and the advantages and disadvantages of the design as a booklet design are discussed.

The literature on experimental design is historically well established and many modern textbooks on this topic (e.g., Giesbrecht & Gumpertz, 2004; Kuehl, 2000) continue to cite traditional sources such as Cochran and Cox (1957) or Kirk (1968). The foundations of a theory of experimental design were laid by Ronald A. Fisher at the beginning of the last century. Experimental design was originally mainly concerned with problems in agriculture, following the primary objectives of (a) the reduction of experimental error by means of local control of experimental conditions, (b) the precise estimation of experimental error by means of replication, and (c) the valid interpretation of the experimental error estimate by means of randomization. Experimental designs are well suited to be used as booklet designs mainly due to their capability to control for unwanted sources of variation that impact item and person parameter estimates and to formalize the efficiency of their estimation.

Basic Terminology

The basic elements in an experiment are known as *experimental units*, the measurements that are taken on the experimental units are known as the *outcomes of interest*, and the interventions that are assigned to the experimental units are known as *treatments*. The term experimental unit denotes a single object or a group of objects to which a treatment is applied in a single trial of an experiment. When the same treatment is applied to multiple experimental units, one speaks of *replication*.

In order to properly estimate *treatment effects* on the outcomes of interest, it is usually necessary to *randomize* the assignment of treatments to experimental units. Randomization is applied to ensure that the systematic effects of all potential confounding factors on the outcomes are identical, on average, for the different treatment levels so that differences in the outcomes of interest can be solely attributed to the variation in the treatments. Quite simply, then, the plan by which the treatments are assigned to the experimental units, given all of the above considerations, is known as the *experimental design.*

A booklet design can be seen as a special case of an experimental design. Within a booklet design, items, testlets, or clusters are the treatments that are assigned to booklets in a systematic way; most frequently, items and testlets are assigned to clusters that are, in turn, assigned to booklets. Thus, in the following sections, we refer to clusters as the treatments in booklet design and only differentiate between items, testlets, and clusters where it is necessary. The frequency with which a cluster is repeated within the complete set of booklets is its replication. The students or groups of students to whom the booklets are given are the experimental units. Finally, the responses given to the items, testlets, or clusters are the outcomes of interest.

Complete Permutation Designs (CPDs)

If unwanted effects of one or more confounding variables on parameter estimates are to be avoided in an experimental design context, one can completely control for the confounding variables by means of complete permutation of treatments. As stated earlier, prominent sources of unwanted variation that should be controlled for in large-scale assessments of student achievement stem from the use of multiple booklets, from the existence of position effects, and from the existence of carryover effects. In CPDs, the order of the treatments is permuted, meaning that every order of clusters appears exactly once in the set of booklets. A simple example with three clusters, represented by the numbers 1, 2, and 3, with two positions per booklet is shown in Table 1. There are three possible ways of selecting two clusters out of the three (1 and 2, 1 and 3, 2 and 3) as well as two ways to arrange the two selected clusters across the two positions. Thus, a total of six different booklets is necessary to cover all potential cluster pairings and permutations. As a result, every cluster of items appears exactly four times, and every order of the clusters appears exactly once. Thus, variations that may be introduced by using a smaller number of booklets as well as by position and carryover effects between the clusters are completely controlled for.

The construction of booklet designs by complete permutation is simple but only practically feasible for small numbers of treatments (i.e., items, testlets, or clusters). Otherwise, such a design would result in a prohibitively large numbers of booklets. Mathematically, with t treatments and k positions per booklet, there are a total of $\binom{t}{k}$ ways in which the treatments could be distributed across the positions. If all possible positions have to be permuted within the designs to further control for position effects, the number of booklets b grows substantially to $b = \binom{t}{k} \cdot k!$.

For example, consider the PISA 2006 study where 13 clusters and four positions in each booklet were used. In this case, one would have needed $\binom{13}{4} = 715$ booklets to cover all possible combinations of clusters in booklets and $715 \cdot 4! = 715 \cdot 12 = 17{,}160$ booklets to completely control for position effects using a CPD. As is the case whenever clusters are assigned to positions using any design, position effects in a CPD can only be controlled for between clusters and not within clusters, because item order within a cluster is not affected by the permutation of the clusters themselves.

Table 1. Booklet Design Based on a Complete Permutation Design

	Booklet					
Position	**1**	**2**	**3**	**4**	**5**	**6**
1	1	2	1	2	3	3
2	2	3	3	1	2	1

Note. The design is based on three clusters, six booklets, and two positions per booklet.

Summing up, booklet designs based on complete permutations have the advantage that they can be constructed easily and that they are capable of controlling for unwanted effects of confounding variables on parameter estimates. The striking disadvantage of those designs, making them infeasible for nearly every large-scale assessment, is that they are only applicable for very small numbers of items, testlets, or clusters. As a result, alternative experimental designs need to be considered for large-scale assessments of student achievement. Specifically, incomplete designs with blocking factors are better suited to be used as booklet designs because they can also control for different sources of unwanted variation but need relatively few booklets. These designs are described in the following sections.

Incomplete Block Designs (ICBDs)

Blocking is one of the most fundamental techniques in experimental design. The basic idea of blocking lies in the removal of unwanted variability in the outcomes of interest. This is done by dividing the collection of experimental units into homogeneous subsets, which are the *blocks*, and randomly applying the treatments to these homogenous subsets. This procedure removes the effect of extraneous sources of variability incorporated into the blocking factor on the parameter estimates of interest.

For booklet design, this implies that factors that may have an unwanted effect on item and person parameter estimates should be included as blocking factors. The two most common effects that researchers would like to eliminate in large-scale assessments are biases due to booklet and position effects. The booklet number and the position number can be viewed as blocking factors because booklets are connected physical units and generic position numbers are naturally identical across booklets. That is, the random assignment of items, testlets, or clusters from an item pool to different positions in different booklets is akin to the random assignment of treatments to experimental units within a two-way blocking structure with booklet and position number as the blocking factors. In complete block designs, every block accommodates the full set of treatments. In terms of booklet design this means that every booklet contains all clusters. Theoretically, *latin square designs* (e.g., Giesbrecht & Gumpertz, 2004)—as special cases of complete block designs—are a good solution to control for booklet and position effects. These are designs with multiple blocking factors such that the number of levels of each factor and the number of treatments are the same. For large-scale assessments, this would imply that the same number of booklets, booklet positions, and clusters need to be used.

However, booklet designs in large-scale assessments of student achievement are typically used with the objective of *not* presenting all clusters to all respondents in the first place. Hence, ICBDs are frequently used as booklet designs. Block designs are called incomplete if the number of treatments per block (i.e., clusters per booklet) is smaller than the overall number of treatments. These designs are well suited as booklet designs because even large numbers of clusters can be assigned to a booklet given a restricted number of positions.

There is a large variety of ICBDs. In the next sections, three types of designs that can be used well as booklet designs are introduced. First, BIBDs are introduced, then YSDs are described, which are a special type of BIBDs and, finally, RTDs are presented.

Table 2. Booklet Design Based on a Balanced Incomplete Block Design

	Booklet						
Position	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	1	2	3	4	1	2	1
2	2	3	4	5	5	6	3
3	4	5	6	7	6	7	7

Note. The design is based on $t = 7$ clusters, $b = 7$ booklets, $r = 3$ occurrences of each cluster, $k = 3$ positions within each booklet, and $\lambda = 1$ occurrences of each cluster pair.

Balanced incomplete block designs (BIBDs). BIBDs are a very large and, possibly, the most important class of ICBDs. In terms of booklet designs, a BIBD is a specific type of an incomplete design that satisfies the following conditions:

1. Every cluster (t) occurs at most once in a booklet (b).
2. Every cluster appears equally often (r) across all booklets.
3. Every booklet is of identical length, containing the same number of clusters (k).
4. Every pair of clusters occurs together in booklets with equal frequency (λ).

The constants t, b, r, k, and λ are called the parameters of the design and characterize the ICBD. For example, the design shown in Table 2 ($t = 7, b = 7, r = 3, k = 3, \lambda = 1$) incorporates seven clusters in three positions across seven booklets with each cluster appearing three times and each pair of clusters appearing together only once.

The design depicted in Table 2 seems practically well suited as a booklet design. The number of booklets is not too large, they are of equal length, and no cluster appears more than once within a single booklet. Furthermore, because all clusters appear with equal frequency in the set of booklets, it is likely that item parameters will be estimated with similar efficiency if all booklets are administered to an equal number of students. An easy way to present all booklets to an equal number of students lies in *spiraling* them across the students, both within and across classrooms. Starting with a first classroom of j students, student 1 is given booklet 1, student 2 is given booklet 2, and so on. If the classroom size j is larger than the number of booklets b, student $j = b + 1$ is presented booklet 1 again until a booklet is assigned to every student in the classroom. In the next classroom, the sequence of booklets is continued. If, for example, student j of the first classroom was given booklet 4, then booklet 5 is given to the first student of the next classroom. Using this simple procedure every booklet is randomly assigned to an (nearly) equal number of students.

The fact that in the design of Table 2 every cluster appears exactly once in conjunction with any other cluster also leads to a homogenous linking between the items. Generally, a booklet design that incorporates just enough clusters to ensure a reliable linkage while maintaining adequate content coverage facilitates the efficient estimation of item and person parameters (see, e.g., the 2007 special issue of the *Journal of Educational Measurement*). Despite these desirable features, a limitation of BIBDs is also apparent from

Table 2. The design does not control for position effects because only one blocking factor is used. For example, the first cluster is always presented in the first position of a booklet and the seventh cluster is always presented at the last position. Obviously, clusters and the position in the booklet are confounded. In a large-scale assessment of student achievement, this may be problematic because the percentage of correct answers tends to decrease with the positions, resulting in biased item parameter estimates. In the example in Table 2, the estimated item difficulties for the items entailed in cluster 10, which is only presented in the last position, will be too high (meaning the items appear too difficult).

Furthermore, not all design parameter combinations are possible for BIBDs. Even if the parameters of a possible BIBD are known, the construction of BIBDs is often not a trivial task and still represents an object of statistical research. Simple BIBD construction techniques can be found in Giesbrecht and Gumpertz (2004). However, because the construction of BIBDs quickly gets very complicated if the number of clusters is more than a few, it is much more purposeful to use designs given by means of tables in standard textbooks of experimental (Cochran & Cox, 1957) and combinatorial design (Colbourn & Dinitz, 1996; Stinson, 2004). These designs can readily be utilized as booklet designs. Doing so, clusters should be randomly assigned to the treatment numbers.

Taken together, BIBDs are well suited to be used as booklet designs for a variety of assessment situations, although the general definition of BIBDs may comprise some disadvantages for specific assessment situations where position effects have to be expected. One striking advantage of BIBDs is their efficiency compared to designs based on complete permutation of the item order. For the case with seven clusters and three positions used as an example in Table 2, complete permutation of the clusters across the booklets with a CPD would produce 35 booklets because there are 35 ways in which three clusters can be selected out of seven. This would be practically infeasible to implement, whereas a BIBD with seven booklets can be implemented. Furthermore, because every cluster has equal replications within a BIBD, the resulting item parameters are likely to be estimated with similar efficiency. Another advantage stems from the fact that the frequency of each pairing of two clusters is held constant in the design leading to a robust linkage across booklets. The disadvantages of BIBDs are that they only exist for some combinations of the design parameters, and that they do not necessarily control for position and carryover effects. To control for position effects, a second blocking factor has to be incorporated, ideally without the restrictions that a latin square design imposes. This can be achieved with YSDs, which are the focus of the next section.

Youden square designs (YSDs). YSDs were introduced by Youden (1937, 1940) to provide efficient experimental designs for biological research. A detailed description of Youden's early greenhouse experiments that led to the development of YSDs can be found in Preece (1990). YSDs are a special case of BIBDs that incorporate the four conditions of a BIBD but impose the additional condition that every cluster has to appear in each position with equal frequency. This additional condition implies that the number of booklets equals the number of clusters (i.e., $b = t$), and that the frequency with which clusters appear in the set of booklets equals the number of positions in the booklets (i.e., $r = k$).

Table 3. Booklet Design Based on a Youden Square Design

	Booklet						
Position	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	1	2	3	4	5	6	7
2	2	3	4	5	6	7	1
3	4	5	6	7	1	2	3

Note. The design is based on $t = 7$ clusters, $b = 7$ booklets, $r = 3$ occurrences of each cluster, $k = 3$ positions within each booklet, and $\lambda = 1$ occurrence of each cluster pair.

Because every cluster appears in every position with equal frequency, position effects are controlled for.

YSDs are frequently used as booklet designs in large-scale assessments of student achievement such as PISA 2003 and PISA 2006 as well as NAEP. While NAEP uses YSDs to construct focused booklets with items from only one subject domain, in PISA a YSD is used to construct a design that mixes subject domains across booklets. Table 3 shows a YSD with $t = 7, b = 7, r = 3, k = 3$, and $\lambda = 1$. Like in the BIBD in Table 2, the seven clusters are distributed in a way that each cluster appears three times and each pair of clusters appears together exactly once. But in distinction to the design in Table 2, the design in Table 3 has the additional feature that each cluster appears exactly once at each position. For example, cluster one is presented at the first position in booklet one, at the second position in booklet seven, and at the last position in booklet five.

Obviously, the YSD presented in Table 3 looks rectangular rather than square, as the name implies. This is due solely to the way in which the design is presented. Originally, the name YSD referred to another form of representation, namely as a $t \times t$ matrix with each cell being either empty or containing the position for the cluster in the particular booklet. We use the rectangular representation here, because it is more compact and mimics the representation of the other designs discussed previously.

Because in YSDs each cluster appears in each position within a booklet exactly once, position effects are controlled for. Despite this advantage compared to a BIBD, a YSD can still not control for carryover effects. Moreover, the advantageous properties of YSDs vis-à-vis alternative designs come at a hefty price of restrictiveness as YSDs only exist for a few combinations of the design parameters. One way to circumvent the availability of a YSD for a particular ideal assessment design scenario is, of course, to choose a YSD design beforehand and arrange the item pool and the assessment characteristics according to the design. When this is not possible, some items from the pool may have to be excluded from consideration in designing the booklets or newly constructed items may have to be added to fill in gaps in the YSD that is chosen after the item pool was developed.

The construction of YSDs is more difficult than the construction of BIBDs. Therefore, for YSDs it is even more necessary than for BIBDs to use readymade designs provided in textbooks. For example, YSDs with $t \leq 91$ are shown in Cochran and Cox (1957). These designs are very useful for the common case where clusters are used to construct booklets. In cases where it seems desirable to assign single items

to booklets (Frey, Carstensen, & Hartig, 2006), larger YSDs are needed. To our knowledge, unfortunately, no complete table of YSDs with $t > 91$ exists.[2]

Summing up, YSDs have the same advantages when used as booklet designs as BIBDs and are thus not repeated here. Furthermore, they not only control for one but for two sources of unwanted variation. This feature can be utilized to control for booklet and position effects. The disadvantages of YSDs are that they only exist for certain parameter combinations, and that they do not control for bias introduced by carryover effects. If carryover effects have to be assumed, RTDs that are described in the following section may be used as booklet designs.

Repeated treatment designs (RTDs). Experimental design distinguishes between strongly balanced and minimally balanced RTDs. In a strongly balanced RTD every treatment follows every other treatment—including itself—an equal number of times. Because clusters cannot appear in booklets multiple times, a strongly balanced RTD is not applicable for the context of booklet design. In a minimally balanced RTD, this structural characteristic is relaxed and every treatment only follows every other treatment an equal number of times, which makes them suitable as booklet designs. RTDs with an equal number of clusters, booklets, and positions are especially useful as booklet designs because they can be constructed easily. The characteristics of an RTD for six clusters, six booklets, and six positions that controls for position effects as well as for first-order carryover effects are demonstrated by the example given in Table 4.

Because in the example every cluster appears in every position exactly once, position effects are controlled for. Furthermore, every cluster is followed by every other cluster exactly once so that first-order carryover effects are also controlled for. However, higher-order carryover effects of individual clusters on clusters presented later on in the booklet are not controlled for by an RTD, similar to a YSD or a general BIBD. It is obvious that RTDs capable of controlling for higher-order carryover effects would be much more complex so that they mostly cannot be used as booklet designs. Because all clusters are presented within all booklets, RTDs share the technical feature with BIBDs and YSDs that every combination of two clusters appears with equal frequency in the set of booklets (e.g., six times in the example in Table 4).

Table 4. Booklet Design Based on a Repeated Treatment Design

	Booklet					
Position	**1**	**2**	**3**	**4**	**5**	**6**
1	1	2	3	4	5	6
2	6	1	2	3	4	5
3	2	3	4	5	6	1
4	5	6	1	2	3	4
5	3	4	5	6	1	2
6	4	5	6	1	2	3

Note. The design is based on $t = 6$ clusters, $b = 6$ booklets, $r = 6$ occurrences of each cluster, $k = 6$ positions within each booklet, and $\lambda = 6$ occurrences of each cluster pair.

An important difference between RTDs and the designs described in the previous sections is that in RTDs every cluster appears in every booklet. This means that RTDs cannot be used in situations where the testing time for all items exceeds the available testing time. They can, however, be of interest if item order or carryover effects are to be controlled for, or if these effects are to be examined more closely. For example, several studies on item position effects in questionnaires made use of these designs (Hamilton & Shuminsky, 1990; Hartig, Hölzel, & Moosbrugger, 2007; Knowles, 1988; Knowles & Byers, 1996). Because the number of booklets equals the number of item clusters, these designs can only be used if the number of clusters is reasonably small.

The construction of RTDs is relatively easy. In the following, an algorithm to construct RTDs with an even number of treatments is described; a slightly more complex algorithm for odd numbers of treatments can be found in Giesbrecht and Gumpertz (2004). First, a *generating column* is generated by writing down the symbols 1, 2, 3, . . . , $t/2$ in the odd-numbered positions and $t, t-1, \ldots, t/2+1$ in the even-numbered positions. The generating column serves as the first column of the design from which the other columns are cyclically generated. This means by going right, for each column, the value is increased by 1; if t is reached, the next column will get a 1. The algorithm for the example above is depicted in Table 5. The six columns at the right-hand side of the table constitute the final design.

Taken together, RTDs are well suited as booklet designs in situations with small numbers of clusters, if carryover effects have to be expected. RTDs have the advantage that they control for two sources of variation and can thus be used to avoid unwanted impact of booklet and position effects on the parameter estimates of interest. Furthermore, they control for first-order carryover effects, and can be constructed relatively easily and without the use of computer programs. The main disadvantage of RTDs lies in the fact that every cluster is presented in every booklet and that they need as many booklets as clusters. This is the reason why RTDs are typically not applied in large-scale assessments of student achievement where large item pools are used.

Table 5. Construction of a Repeated Treatment Design

		Column					
Generating Column		**1**	**2**	**3**	**4**	**5**	**6**
1	→	1	2	3	4	5	6
6	→	6	1	2	3	4	5
2	→	2	3	4	5	6	1
5	→	5	6	1	2	3	4
3	→	3	4	5	6	1	2
4	→	4	5	6	1	2	3

Note. The design is based on an even number of $t = 6$ clusters.

Evaluation of Booklet Designs

When a booklet design has to be constructed for a large-scale assessment of student achievement there are usually several candidate designs that seem feasible. In most cases, it is

quite difficult to decide which design has the most desirable attributes and should be used as the final booklet design as each design entails compromises. Thus, a reasonable strategy is to base the decision for or against a design on a mixture of information based on qualitative and quantitative criteria, taking the special characteristics of the assessment situation into account.

Information based on *qualitative criteria* can be obtained by answering questions like the following for any of the candidate designs. Even if most of the questions appear rather trivial perhaps, setting up a catalog and systematically answering them for each candidate design makes it easier to avoid overlooking an important characteristic.

1. Does the design meet all the restrictions of the assessment situation?
   - E.g., is the design applicable from a practical point of view? For example, is the testing time that is allocated to a booklet by the design aligned with the actual available testing time?
   - E.g., does the design control for possible sources of unwanted variation? For example, does the design control for position and/or carryover effects of a specified order?
2. Are all factors that the design controls for sources of unwanted variation? For example, if carryover effects will not be a problem, a design that controls for position effects is not needed.

If any of these questions can be answered with "no", the corresponding design is likely to be problematic and the question of whether other designs are possible must be seriously considered.

Additionally, information based on *quantitative criteria* can be used to evaluate booklet designs. In most cases, the *variance-covariance matrices of the design matrix* and the *D-optimality index* are particularly useful. The variance-covariance matrix of the design matrix shows the degree of association between the factors incorporated into the design. Within booklet designs, these factors are booklet, position, and cluster. To calculate the variance-covariance matrix, the design has to be represented as a design matrix, which is a different representation of the design compared to the tables that have been used so far in this module. Specifically, the design matrix has to contain a separate column for each factor, which results in every cell of the design being represented by a row in the design matrix.

To illustrate, Table 6 shows a small YSD with $t = 3$, $b = 3$, $k = 2$, $r = 2$, $\lambda = 1$, in the format that was used for all designs in this module, while Table 7 shows the same design in the design matrix format that is required to estimate the variance-covariance matrix.

Based on the representation in Table 7, the variance-covariance matrix can easily be calculated with standard statistical software programs. The variance-covariance for the YSD in Tables 6 and 7 is shown in Table 8. Rearranging designs can consume a lot of time and may result in mistakes if the designs have a lot of cells. To avoid this, the computer program mentioned in Footnote 2 can be used to calculate variance-covariance matrices for designs constructed with the computer program itself or for designs that were read in from external ASCII files.

As is typical for such matrices, the elements on the main diagonal of the variance-covariance matrix are the variances of the three design factors and the off-diagonal elements are their covariances. In this example, the variance of the factors booklet and cluster is higher than the variance of the factor position because the former have three levels whereas the latter has only two levels. In other words, one can learn more from this design about the effects of booklets and clusters on item and person parameter estimates than about the effects of cluster position on these estimates.

Which values for the variances and the covariances are desirable cannot be stated generally and has to be determined anew for every study. For example, if the objective of a large-scale assessment of student achievement is to analyze the effects of using various numbers of booklets on parameter estimates, then the variance of the booklet factor should be high. This can be achieved by incorporating many different booklets, because the variance of a factor rises with the number of its levels.

Table 6. Booklet Design Based on a Youden Square Design in Typical Format

	Booklet		
Position	**1**	**2**	**3**
1	1	2	3
2	2	3	1

Note. The design is based on $t = 3$ clusters, $b = 3$ booklets, $r = 2$ occurrences of each cluster, $k = 2$ positions within each booklet, and $\lambda = 1$ occurrences of each cluster pair.

Table 7. Booklet Design Based on a Youden Square Design as a Design Matrix

Booklet	Position	Cluster
1	1	1
1	2	2
2	1	2
2	2	3
3	1	3
3	2	1

Note. The design is based on $t = 3$ clusters, $b = 3$ booklets, $r = 2$ occurrences of each cluster, $k = 2$ positions within each booklet, and $\lambda = 1$ occurrences of each cluster pair.

Table 8. Variance-Covariance Matrix for a Youden Square Design

	Booklet	Position	Cluster
Booklet	.80	.00	.20
Position	.00	.30	.00
Cluster	.20	.00	.80

Note. The analyses refer to a design with $t = 3$ clusters, $b = 3$ booklets, $r = 2$ occurrences of each cluster, $k = 2$ positions within each booklet, and $\lambda = 1$ occurrences of each cluster pair.

In practice, however, large-scale assessments of student achievement are generally not used to examine the effects of design factors on parameter estimates, but rather to control for unwanted sources of variability. Thus, for most large-scale assessments, the covariances are more relevant for the evaluation of booklet designs than the variances. If the covariance between two factors is 0, these factors are said to be orthogonal, meaning they are statistically unrelated by design. In this case, no bias can result from an interaction of the two factors. If the covariance is different from 0, the factors are said to be confounded, meaning they are statistically related by design. The variance introduced by confounded factors cannot be separated in the statistical analysis later on and may lead to bias in parameter estimates. Often, confounded factors are problematic but they do not necessarily have to be. For every assessment situation, the question of whether a calculated covariance is problematic or not must be thoroughly thought over.

A situation where a covariance between the factors position and cluster is not wanted is the case when position effects have to be expected and one objective of the study is the estimation of item parameters. The design shown in Table 6—like all YSDs—is unproblematic in this regard, with a covariance of 0 between cluster and position. The covariance of .20 between booklet and cluster is due to the clusters not being evenly distributed across the booklets. An even distribution is, in fact, impossible in the example because there are more clusters than positions in a booklet. Because the parameter estimates are calculated on the basis of data that are aggregated across booklets in large-scale assessments of student achievement, the confounding of the factors cluster and booklet does not lead to bias in parameter estimates and can therefore be seen as unproblematic. Obviously, when the covariance between booklet and cluster is different to 0, parameter estimates should not be obtained from data drawn from a subset of the booklets. In this case, biased estimates would result.

One way to compare different variance-covariance matrices for a given assessment scenario is to use a particular optimality criterion. For applications of booklet designs in large-scale assessments, the D-optimality index (e.g., Atkinson & Donev, 1992; van der Linden, 2005) is the most important optimality criterion. The D-optimality index is an aggregated measure of the variance-covariance matrix and is calculated by taking the determinant of the design matrix multiplied by the inverse of the design matrix. Thus, the value of the D-optimality index will become large if the values on the main diagonal of the design matrix are large relative to the elements off the main diagonal. The higher the values of the D-optimality index, the more independent the factors of the design are.

In summary, booklet designs can be evaluated and compared with each other by a combination of qualitative and quantitative criteria. No standard procedure or standardized cut-off values for variances or covariances are available, as assessment objectives and design constraints differ widely between different large-scale assessments of student achievement. Hence, we advise to ask and answer critical qualitative questions about the desired classes of inferences about students and assessment characteristics in a first step to check the extent to which any candidate design meets the requirements of the current assessment situation. We further propose to use quantitative criteria such as the variance-covariance matrices and the D-optimality indices for different candidate designs to examine their relative suitability in a second step.

Conclusion

This module has shown that the construction of a booklet design for a large-scale assessment of student achievement is a complex and nontrivial task. Constructing a booklet design is fundamentally the task of finding a design structure that specifies the assignment of clusters to booklets in order to meet the objectives of a study, given a set of constraints posed by the assessment situation. The general objective of large-scale assessments of student achievement lies in validly reporting student proficiencies, mostly at the level of groups. This general objective can only be reached if multiple statistical objectives are fulfilled. For most assessment situations, the statistical objectives are to obtain unbiased and efficient item and/or person parameter estimates, requiring that unwanted sources of variation of these parameter estimates, such as variation due to position and carryover effects, are controlled for as much as possible.

This can be achieved primarily in the early phases of test construction and field testing by extensive training of item writers that leads to well-targeted items with minimal degrees of nuisance dependency, by keeping testing sessions reasonably short, and by avoiding the creation of item clusters and cluster arrangements that change response processes in an unwanted manner. After the item pool has been developed, a suitable booklet design needs to be constructed to statistically control for possible remaining source of unwanted variation as much as possible. Finally, the booklets need to be presented to a sufficiently large, random, and representative sample of the target population of students.

Due to the complexity and interdependency of these tasks, no straightforward general design algorithm can be utilized for the construction of booklet designs. Because large-scale assessments of student achievement can differ widely with regard to their substantive and statistical objectives as well as their practical implementation constraints, it is typically necessary to construct a booklet design anew for each study. To reduce the effort associated with booklet design construction, systematically constructed designs from the combinatorics and experimental design literature can be used to alleviate the burden for the assessment developers. In this module, we reviewed several key design types that fit this mold, which included CPDs, BIBDs, YSDs, and RTDs. Among these designs, BIBDs and YSDs in particular are attractive design types that can and have been used successfully in large-scale assessments of student achievement.

Although all of these designs statistically control for one or more unwanted sources of variation, it is important to keep in mind that unwanted effects on the parameter estimates of interest are strictly speaking only completely "removed" if the mean of the effects equals 0 across the levels of the respective factor (e.g., across the positions in a set of booklets). If this is not the case, the bias in item parameter estimates has to be acknowledged when inferences are drawn from the test scores. However, in most cases, the bias caused by unwanted effects on the parameter estimates of interest in balanced designs is rather small compared to the bias resulting from unbalanced designs.

One methodological lesson that can be learned from all of this is that it is worthwhile to align the development of the item pool not merely with a general table of specification for the assessment but also with the booklet design that will be subsequently implemented to collect the data. Because all decisions about students will be made on the basis of the collected data, which critically depend on the quality of the booklet design, developing suitable booklet designs should be of primary concern to assessment developers. More important, latent variable models, including IRT models, mostly cannot make up for weaknesses in the booklet design once the data have been collected.

Once suitable designs have been constructed and implemented for a particular study, it may seem to be relatively easy to modify one or more existing candidate designs to accommodate the objectives of a particular assessment. These modifications have to be made with caution, however, because they will distort the statistical properties of the original design to some degree and minor changes may lead to rather large unwanted statistical problems. Thus, the extent to which changes to the design structure result in severe problems or can be tolerated in an assessment situation should be evaluated thoroughly. This can be done with the combination of qualitative and quantitative criteria that are proposed in this module for the evaluation and comparison of different booklet designs. Both types of criteria supply complementary information about whether the designs under consideration introduce unwanted variability in parameter estimates and whether they are generally well suited as booklet designs for the current assessment situation.

It is also worth noting that a majority of booklet designs that are used in current large-scale assessments of student achievement assign clusters, rather than items or testlets, to positions in booklets even though the latter two are selected from the item pool to build the clusters. Although this approach allows quite simple booklet designs with small numbers of booklets to be used, it is not completely unproblematic. Most critically, a suitable booklet design can only ensure that sources of unwanted variation are controlled for between clusters but never within clusters as the cluster construction happens before the cluster assignment. For example, position effects or carryover effects are not controlled for within clusters because the order of items within a cluster is the same for all booklets unless the order is specifically varied. An alternative to the common practice of using clusters would be to regard items or testlets as the smallest possible units and balance their position across booklets. In a simulation study, Frey et al. (2006) showed that assigning single items to booklets instead of clusters results in small but stable advantages regarding bias and efficiency of both item and person parameter estimates. However, assigning items to booklets instead of clusters results in large numbers of booklets, which has several practical drawbacks. As stated earlier in the module, constructing and formatting these booklets consumes a lot of resources and can produce errors if not done by a computer. More critically, a large number of booklets can produce problems regarding the validity of the inferences that are drawn from the test scores. Consider the case where one wants to compare the mean proficiency levels of school classes with each other. In that case, one faces challenges if the number of booklets used is larger than the number of students in a classroom, because in some classrooms no representative sample of items is presented to the students. For example, if a lot of items from one subdomain (e.g., linear algebra) are given to the students of one classroom whereas the students of another classroom get only a few or no items measuring this subdomain, the validity of the interpretation of the difference between the means of the two classrooms as differences in the major content domain (e.g., mathematics) are likely to be problematic as the construct is represented differently for these student groups. To avoid these problems, the number of booklets in the design should be smaller than the number of students in the smallest group for which results will be computed and reported. Furthermore, the booklets should be randomly assigned to students, which can easily be accomplished by spiraling them across the students within classrooms.

One related research question that has not yet been satisfactorily answered in the literature concerns the question of which kinds of booklet design can deal adequately with item pools that measure more than one dimension and lead to a reliable scaling of the response data with multidimensional IRT models. The designs presented in this module can only control for up to two sources of unwanted variation, namely booklet and position. In the case of a multidimensional item pool, dimension can be viewed as an additional blocking factor. If dimension is not balanced by means of the booklet design, the efficiency of the variances and covariances of person parameters is likely to vary between dimensions or between pairs of dimensions, respectively. Unfortunately, balanced designs with three or more blocking factors have not been used as booklet designs in practice so far even though they have been presented in the combinatorics and experimental design literature.

Finally, we want to end the module with a call to action for assessment developers, measurement specialists, and related professionals. It is critical that the booklet designs that have been used in large-scale assessments are more frequently discussed in the educational measurement literature from theoretical, applied, and didactic perspectives. More transparency about the factors that have guided a particular booklet design is clearly needed to lift the construction of booklet design from the realm of diffuse expertise possessed by a selected few to the realm of shared expertise possessed by a discipline. Booklet designs should also have a prominent place in graduate-level training in educational measurement programs, where they are currently frequently neglected. This requires that textbooks or handbooks of measurement devote more space to the structure, properties, construction, and implementation of these designs. Future research needs to investigate how the set of existing booklet designs can be extended to accommodate the real-life complexities of existing item pools and should provide more hands-on tools for assessment developers that aid them in the development of such designs.

Self-Test

1. Explain the primary reason for utilizing booklet designs in large-scale assessments of student achievement.
2. What are two primary statistical objectives that often guide the construction of a booklet design? How are they related to the validity of inferences that are drawn from the test scores?
3. Explain why complete permutation designs are of limited practical use for large-scale assessments.

4. What is the definition of position effects? What are possible reasons for position effects in large-scale assessments of student achievement?
5. What are carryover effects? Give an example for a possible occurrence of carryover effects in large-scale assessments of student achievement.
6. Why is it necessary to control for position effects and carryover effects?
7. Which booklet designs statistically control for position and first-order carryover effects? Why is this the case?
8. Construct a feasible booklet design for the assessment situation described below. Note that you will need a textbook on experimental design like Cochran and Cox (1957) or Giesbrecht and Gumpertz (2004) to inspect tables of candidate designs. The statistical objectives of the assessment are to estimate item and person parameters. The study aims at reporting at the classroom level and at higher levels. The assessment is constrained by the following requirements:
   (a) The item pool contains 130 items measuring mathematical literacy.
   (b) Each item takes approximately 2 minutes to complete.
   (c) The available testing time is 90 minutes.
   (d) The available sample size will be roughly $N = 5{,}000$ students of grade 8.
9. What is a disadvantage of Youden square designs?
10. Construct a minimal balanced repeated treatment design for $t = 4$ clusters. Please calculate the variance-covariance matrix and evaluate the design regarding confounding of the design factors cluster, booklet, and position.

Answers to Self-Test

1. The primary reason is that there are generally more items available in an item pool than any single student can answer within the testing time allotted. To achieve suitable content domain coverage while controlling for unwanted effects on parameter estimates like position effects or carryover effects, the items, testlets, or clusters have to be systematically assigned to different booklets for which a suitable booklet design is required.
2. The two primary statistical desiderata are to estimate unbiased and efficient item parameters as well as to estimate unbiased and efficient person parameters. Estimation of item parameters is typically of primary concern in early phases of a large-scale assessment of student achievement when the psychometric properties of the instruments are established. After a sound psychometric test is established, distributions of person proficiencies are calculated for reporting. When the estimated person parameters are unbiased, inferences based on them are more likely to be appropriate and when they are efficient, inferences will also be reasonably precise. This allows to draw valid inferences from the test scores and to test statistical hypotheses with high statistical power.
3. Even though complete permutation designs control for all unwanted sources of variation, they require an impractical number of booklets to be used.
4. Position effects are effects of item position within a booklet on item difficulty, for example, if correct answers are given more frequently if an item is presented in a cluster at the beginning of a booklet rather than at the end of a booklet. Position effects may be caused by fatigue or a reduction in test-taking motivation with increasing test length or just by a lack of time on the part of the student.
5. Carryover effects are effects of previously presented clusters on response probabilities of subsequent clusters. Carryover effects are a special kind of position effects. When the concern is about the effect of a particular cluster on the following one, this is referred to as a first-order carryover effect whereas effects across more positions are referred to as higher-order carryover effects. For example, a first-order carryover effect is present if a particular algebra cluster is presented after the student has already worked on a number of similar algebra clusters, and he or she will find it easier than if the same cluster is presented after working on a number of geometry clusters or even distinct clusters from a different content domain such as reading.
6. IRT-based scaling models assume invariance of item parameters across multiple student groups or design conditions and local independence of response probabilities for students with identical proficiencies across items. Position effects violate the assumption of invariance, because item difficulty varies with respect to the position an item is presented in a booklet. If unaccounted for, variations in item parameter estimates due to position effects can lead to severe validity problems if the item difficulties are used to give meaning to sections of a scale, which can transform the problem from a measurement problem to a reporting problem. Carryover effects violate the assumption of local independence and can also cause severe validity problems.

Table 9. Booklet Design Based on a Youden Square Design

	Booklet												
Position	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	1	2	3	4	5	6	7	8	9	10	11	12	13
2	2	3	4	5	6	7	8	9	10	11	12	13	1
	Break (10 Minutes)												
3	4	5	6	7	8	9	10	11	12	13	1	2	3
4	10	11	12	13	1	2	3	4	5	6	7	8	9

Note. The design is based on $t = 13$ clusters, $b = 13$ booklets, $r = 4$ occurrences of each cluster, $k = 4$ positions within each booklet, and $\lambda = 1$ occurrence of each cluster pair.

Table 10. Booklet Design Based on a Repeated Treatment Design

	Booklet			
Position	**1**	**2**	**3**	**4**
1	1	2	3	4
2	4	1	2	3
3	2	3	4	1
4	3	4	1	2

Note. The design is based on $t = 4$ clusters, $b = 4$ booklets, $r = 4$ occurrences of each cluster, $k = 4$ positions within each booklet, and $\lambda = 4$ occurrences of each cluster pair.

7. Position effects and first-order carryover effects can be controlled for by minimally balanced repeated treatment designs. Because in repeated treatment designs every cluster appears in every position exactly once, position effects are controlled for. Furthermore, every cluster is followed by every other cluster exactly once so that first-order carryover effects are also controlled for.
8. The size of the item pool does not allow presenting all items to each individual student. Because the testing time is rather long for eighth graders, position effects may occur due to fatigue effects. However, carryover effects are not likely to be a problem because only one content domain is measured. A feasible booklet design for the described assessment situation would be a Youden square design with $t = 13$ clusters entailing 10 items each, $b = 13$ booklets with $k = 4$ positions each, each cluster appearing $r = 4$ times, and each pair of clusters appearing $\lambda = 1$ times. To reduce possible position effects due to fatigue on the side of the students, a break can be included in the middle of the testing session (see Table 9).
9. Youden square designs only exist for a few combinations of the design parameters (number of clusters, booklets, and positions in booklets).
10. One possible repeated treatment design is given in Table 10. Because all values on the main diagonal of the variance-covariance matrix are 1, and all other values are 0, the design factors are not confounded. Thus, the design controls for position effects and first-order carryover effects.

Notes

[1]In the following text, the term *control* is used if we refer to statistical control by averaging out unwanted variability in parameter estimates. [2]To solve this shortcoming, we are developing a computer program that constructs YSDs for given combinations of design parameters. The program is written in Java and has a user-friendly point-and-click interface. Currently, the program is capable of finding about 60% of the existing YSDs. Readers interested in the program may contact the first author for a copy.

References

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). NAEP 1998 technical report. Jessup, MD: Education Publications Center. (http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2001509, accessed September 15, 2008.)

Atkinson, A. C., & Donev, A. N. (1992). *Optimum experimental designs*. Oxford, UK: Clarendon.

Beaton, A. E., & Zwick, R. (1992). Overview of the national assessment of educational progress. *Journal of Educational Statistics*, *17*, 95–109.

Brennan, R. L. (Ed.) (2007). *Educational measurement* (4th ed.). Westport, CT: Greenwood Publishing Group, Inc.

Briggs, D. (2009, April). *Measuring and evaluating change in student achievement: A conversation about technical and conceptual issues*. Symposium presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Childs, R. A., & Jaciw, A. P. (2003). Matrix sampling of items in large-scale assessments. *Practical Assessment, Research & Evaluation*, *8*(16), 1–11. (http://PAREonline.net/getvn.asp?v=8&n=16, accessed September 15, 2008).

Cochran, W. G., & Cox, G. M. (1957). *Experimental design*. New York: John Wiley & Sons.

Colbourn, C. J., & Dinitz, J. H. (Eds.) (1996). *The CRC handbook of combinatorial designs*. Boca Raton, FL: CRC Press.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

de Boeck, P., & Wilson, M. (2004). *Explanatory item response modeling*. New York: Springer.

Downing, S. M., & Haladyna, T. M. (Eds.) (2006). *Handbook of test development*. Mahwah, NJ: Erlbaum.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Frey, A., Carstensen, C. H., & Hartig, J. (2006). *BIB-designs in large scale assessments: Should single items or clusters of items used to assemble booklets?* Paper presented at the 71st Annual Meeting of the Psychometric Society (IMPS), Montreal, Canada.

Frey, A., & Seitz, N. N. (in press). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*.

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

Hamilton, J. C., & Shuminsky, T. R. (1990). Self-awareness mediates the relationship between serial position and item reliability. *Journal of Personality and Social Psychology*, *59*, 1301–1307.

Hartig, J., Hölzel, B., & Moosbrugger, H. (2007). A confirmatory analysis of item reliability trends (CAIRT): Differentiating true score and error variance in the analysis of item context effects. *Multivariate Behavioral Research Methods*, *42*, 157–183.

Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Pacific Grove, CA: Wadsworth.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, *55*, 312–320.

Knowles, E. S., & Byers, B. (1996). Reliability shifts in measurement reactivity: Driven by content engagement or self-engagement? *Journal of Personality and Social Psychology*, *70*, 1080–1090.

Kolen, M. L., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.

Kuehl, R. O. (2000). *Design of experiments: Statistical principles of research design and analysis* (2nd ed.). Pacific Grove, CA: Duxbury/Thomson Learning.

Le, L. T. (in press). Effects of item positions on their difficulty and discrimination: A study in PISA science data across test language and countries. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics*. Tokyo: Universal Academic Press.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. (http://timss.bc.edu/pirls2006/tech_rpt.html, accessed September 15, 2008.)

McDonald, R. F. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Muthén, B. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 81–117.

OECD (2005). *PISA 2003 technical report*. Paris: OECD. (http://www.pisa.oecd.org/dataoecd/49/60/35188570.pdf, accessed September 15, 2008.)

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

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.) (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. (http://timss.bc.edu/TIMSS2007/techreport.html, accessed May 2, 2009.)

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Preece, D. A. (1990). Fifty years of Youden squares: A review. *Bulletin of the Institute of Mathematics and Its Applications*, *26*, 65–75.

Rao, C. R., & Sinharay, S. (Eds.) (2007). *Handbook of statistics, volume 26: Psychometrics*. Amsterdam: North-Holland.

Rupp, A. A., & Templin, J. (2008). Unique characteristics of cognitive diagnosis models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, *6*, 219–262.

Rupp, A. A., Templin, J., & Henson, R. (in press). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

Rupp, A. A., Vock, M., Harsch, C., & Köller, O. (2008). *Developing standards-based assessment items for English as a first foreign language: Context, processes, and outcomes in Germany*. Münster: Waxmann.

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 429–438). Amsterdam: Elsevier.

Stinson, D. R. (2004). *Combinatorial designs: Construction and analysis*. Berlin: Springer.

Thissen, D., & Wainer, H. (Eds.) (2001). *Test scoring*. Mahwah, NJ: Erlbaum.

Tucker, A. (2006). *Applied combinatorics*. New York: Wiley.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer.

van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing incomplete block designs for large-scale educational assessments. *Applied Psychological Measurement*, *28*, 317–331.

Verschoor, A. J. (2007). *Genetic algorithms for automated test assembly*. Arnhem, The Netherlands: CITO.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in national assessment of educational progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, volume 26: Psychometrics* (pp. 1039–1056). New York: Elsevier.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Yen, W., & Fitzpatrick, A. R. (2006). Item response theory. In R. J. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: Praeger.

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.

Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement*. Princeton, NJ: Educational Testing Service.