Psychological Test and Assessment Modeling, Volume 54, 2012 (4), 366-396

Principles and procedures of considering item sequence effects in the development of calibrated item pools: Conceptual analysis and empirical illustration

Safir Yousfi[1] *& Hendryk F. Böhme*[2]

Abstract

Item responses can be context-sensitive. Consequently, composing test forms flexibly from a calibrated item pool requires considering potential context effects. This paper focuses on context effects that are related to the item sequence. It is argued that sequence effects are not necessarily a violation of item response theory but that item response theory offers a powerful tool to analyze them. If sequence effects are substantial, test forms cannot be composed flexibly on the basis of a calibrated item pool, which precludes applications like computerized adaptive testing. In contrast, minor sequence effects do not thwart applications of calibrated item pools. Strategies to minimize the detrimental impact of sequence effects on item parameters are discussed and integrated into a nomenclature that addresses the major features of item calibration designs. An example of an item calibration design demonstrates how this nomenclature can guide the process of developing a calibrated item pool.

Key words: context effects, sequence effects, item calibration design, item pool development

[1] *Correspondence concerning this article should be addressed to:* Dr. Safir Yousfi, Psychological Research and Development, German Federal Employment Agency, Regensburger Strasse 104, 90478 Nuremberg, Germany; email: safir.yousfi@arbeitsagentur.de

[2] Department of Psychological Research and Development, German Federal Employment Agency

Practical applications of educational and psychological testing often require the administration of different test forms. For example, in low-stakes large-scale testing programs it would lead to an excessive amount of testing time if all items that are needed to cover the respective content domain would be applied to all test takers. Sometimes the selection of items must be tailored to the level of ability, competence, grade, or age in order to be suited for each test taker[3]. In high-stakes contexts concerns about the confidentiality of the items can require the administration of different test forms.

This in turn requires taking care that the expected test score of a test taker does not depend on the test form. Classical solutions to this problem entail administering each test form to large samples of test takers before the operational use of the test. Then an equating function on the test scores is established that eliminates or at least minimizes the dependency of the test score from the administered test form.

In contrast, methods of parametric item response theory (IRT) do not necessarily require administering all test forms in an equating study. If item parameters are available that do not depend on the composition of a test form, then the person parameter estimates of all test forms refer to the same scale. Consequently, test scores that are derived from these person parameter estimates can be interpreted without taking into account which test form was administered. Computerized adaptive testing (CAT) exploits this feature by estimating person parameters after each item response and by optimizing the selection of the remaining items with respect to some criteria (e.g., measurement accuracy or item security). By this way each test taker receives an (almost) unique test form that is tailored to his response behavior and possibly to background variables, too (Eggen & Verschoor, 2006).

It is well known that CAT, like any other method that flexibly composes test forms from a calibrated item pool, relies on the assumption that item parameters do not depend on the test form (*item fungibility*; cf. Wainer & Mislevy, 2000). Substantial context effects thwart any strategy of test development that is based on calibrated item pools. However, Wainer, Bradlow, and Wang (2007) claim that the absence of context effects is usually only postulated in operational computerized adaptive tests without being tested. On the other hand, there is evidence that the response to an item can depend on context factors like the number and the content of preceding items (Wainer et al., 2007; Hohensinn, Kubinger, Reif, Holocher-Ertl, Khorramdel, & Frebort, 2008). Hence, neglecting is possibly not the best strategy of dealing with context effects although it seems to be the most common one. The main purpose of this paper is to encourage test developers to follow other strategies to deal with context effects.

At first we like to clarify relevant terms of test development and IRT. In contrast to converse statements in the literature it is emphasized that context effects are not necessarily a violation of IRT. Instead, context effects, in general, and sequence effects, in particular, can be analyzed by IRT methods in order to ensure that they do not pose a serious threat

[3] Items might not be suited if they refer to curricula of higher grades or if they require skills (like extracting a root) that cannot be expected in the respective age. Responses to items which are too easy or too difficult are not informative.

to the validity of the test scores. Options for the design of a calibration study will be compared with respect to the way they control for sequence effects. Finally, it will be shown by means of a practical example how the outlined conceptual framework can guide the development of a calibrated item pool.

This paper focuses on strategies and concepts to deal with sequence effects in the process of developing a calibrated item pool. Nevertheless, a great part applies to context effects in general. A formalized IRT approach that is compatible with all the concepts that are discussed in this paper was developed by von Davier and von Davier (2007, 2011) although their work was not targeted on considering context effects.

Basic terms

An educational or psychological *test* is a structured situation that is presented to a test taker in order to evoke diagnostically relevant behavior. A *test form* is defined by all factors which are devised by the test developer for a single administration of an educational or psychological test[4]. Every combination of factors that is potentially realized during a test administration defines a different test form if these factors are varied systematically and deliberately. One example is a test for which a paper-and-pencil version and a computer form have been developed. Factors that vary only haphazardly like the location where the test is administered do not constitute different test forms[5].

In accordance with Gonzalez and Rutkowski (2009) we define an *item* as the most basic unit of a psychological or educational test. It is usually an individual task administered to a test taker. In order to aggregate the responses to the items to the final test result, the test taker's response behavior to the respective tasks is usually mapped onto a numerical scale. This process is called item scoring.

The *item pool* is the set of items that is available for assembling test forms. Frequently, test forms differ only in the selection and order of the items from the item pool. This case is the focus of our paper.

A *testlet* is a subset of the item pool that is administered to test takers in a specified sequence at a stretch. In general, a test form is not constituted by a single testlet, but by a sequence of testlets (and/or items). The items of a testlet might refer to the same stimulus (e.g., a reading passage). However, this is no requirement. Some calibration designs split the item pool into testlets. Other prevalent synonyms for testlet are cluster, packet, and block (Wainer & Mislevy, 2000; van der Linden, Veldkamp, & Carlson, 2004; Gonzalez & Rutkowski, 2009). However, we like to discourage using the term block in this context because it has different meanings in the field of experimental design (e.g., Chochran &

[4] The term booklet is often used as a synonym for test form. However, this could be misleading if not all of the test forms are presented in a paper-and-pencil mode.

[5] Sometimes it is not easy to differentiate test forms. For example, if test order is varied systematically (Khorramdel & Frebort, 2011) test forms might differ only in their position in the test battery. However, if test order is only a haphazard by-product of applying different tests in a session, then the position of the test is not a defining characteristic of the test form.

Cox, 1957, or Giesbrecht & Gumpertz, 2004) that inspired most of the methods that were developed for calibration studies. If concepts of experimental design are adapted to test design, block is often used as a synonym for test form (van der Linden et al., 2004; Frey, Hartig & Rupp, 2009).

Linking test forms by means of parametric Item Response Theory

Parametric models of IRT (e.g., van der Linden & Hambleton, 1997, for an overview) express the probability distribution of the item scores of a test form as a function[6] of *item* and *person parameters*. The item parameters are assumed to be constant within each subpopulation of test takers. In order to get empirical estimates of the item parameters it is necessary to conduct a calibration study that administers the test form to a sample of each subpopulation. Models that allow for parameters (referring to items or to the trait-distribution of test takers) that vary *across* subpopulations of test takers are referred to as multi-group models if the variable that characterizes the subpopulations is an observed variable (Bock & Zimowski, 1997) and as mixture models if the respective variable that characterizes the subpopulations is a latent variable (Rost, 1997). *Measurement invariance* refers to a situation in which the item parameters of a test form do not vary across subpopulations. However, even in case of measurement invariance the item parameters are not unambiguously defined. Most IRT models require some arbitrary choices to fix the scale of the person and item parameters because the level of measurement is not absolute[7] (Kolen & Brennan, 2004).

If more than one test form is assembled and an IRT model is established for each test form separately, the question arises to what extent the respective item and person parameters are comparable. The lack of comparability might be attributable to the arbitrary choices that have been made to fix the scale of the item and person parameter space. In this case comparability might be established by a linking transformation that attunes the arbitrary choices to fix the scale. These transformations rely on (implicit or explicit) equality constraints. These constraints can refer to the item parameters of different test forms or to the parameters of the distribution of person parameters in the respective subpopulation. If the equality constraints are formulated for the distribution of person parameters it is necessary to ensure that these parameters refer to the same population of test takers. If the test forms are given to different populations of test takers, consistency of the item parameters might be ensured by equality constraints that refer to the item parameters (Kolen & Brennan, 2004). Most of these linking techniques do not assume complete consistency of the item parameters across test forms but only formulate an equality constraint on a function of the item parameters (e.g., the mean-mean and the mean-var method) or minimize a discrepancy function (e.g., characteristic curve meth-

[6] The respective function of a single item is called *item characteristic function*.

[7] Many prominent models of IRT reach only an interval level of measurement (for the person parameter). In this case different choices to fix the scale of the item and person parameters differ only by a linear transformation.

ods). Only concurrent calibration enforces complete consistency of the item parameters. In any case, equality constraints on item parameters are only justified if the arbitrary choices that have been made to fix the scale are indeed the only reason for the lack of consistency of item parameters across test forms (von Davier & von Davier, 2007, 2011).

Context effects

Differences between item parameters that refer to the same item in different test forms cannot always be attributed to arbitrary choices made to fix the parameter space. The response behavior of test takers to an item needs not to be the same for all test forms in which the item appears. As a consequence the probability distribution of the scores of the item is not independent from the choice of the test form. In this case the inconsistencies of the item parameter estimates cannot be resolved by a linking transformation. Techniques that enforce consistency of item parameters like concurrent calibration (Kolen & Brennan, 2004) are not suited in this case because the equality assumptions that they rely on are violated. Consequently, the parameters of an item depend on the test form, i.e., an item can have as many item parameters as test forms that include the respective item. In the worst case this could even mean that the respective IRT model (i.e., not only the IRT-parameters) that is used to describe the response behavior to this item does not apply to all test forms. Accordingly, an item parameter might not even be defined for some of the test forms. But even if the IRT model applies to all test forms, it is hard to tell which of the item parameters (referring to the different test forms) is the genuine one.

In order to assess the plausibility of equality (or consistency) assumptions of item parameters across test forms it is necessary to address possible reasons for differences. In general, all factors that vary across test forms can be subsumed as context effects[8]. Examples for context effects are the mode of presentation (computer vs. paper-and-pencil) or the instruction that is given. This paper focuses on context effects that are due to the choice and the order of the items of a common item pool for all test forms. These factors will be referred to as *sequence effects*.

Sequence effects

Sequence effects might be differentiated into (1) position effects, (2) order effects, and (3) carry-over effects. (1) Position effects refer to a case in which the difference of the parameters of an item can be attributed to the fact that it appears at different positions in different test forms (i.e., as first item, as second item, etc.). For instance, it is often observed that the first item has a poorer model fit than other items because the test taker is not adjusted to the test situation, yet. Another example for a position effect is that items

[8] Note that this definition of context effects excludes all factors that are not defining characteristics of a test form (like internal states and external circumstances). These factors are usually not considered explicitly, but they are included in the concepts of measurement error and construct-irrelevant variance (Messick, 1995).

at the end of a test appear as more difficult because of fatigue (Hohensinn et al., 2008, Le, 2009). (2) Order effects refer to effects on the item parameters that cannot be attributed to the choice of items but to the order that has been specified after the choice has been made. (3) A carry-over effect refers to a case in which the parameters of an item change due to preceding items. An example for a carry-over effect is the use of the content of an item for solving a subsequent item. A first-order carry-over effect is the effect of the presentation of an item on the item that follows immediately in the respective test form, whereas higher-order carry-over effects refer to items that follow later. For instance, a second-order carry-over effect refers to the item that follows after the next item. Position, order, and carry-over effects are overlapping categories. For instance, sequence effects for an item pool that consists of only two items might be described as position, carry-over, or order effects.

The absence of sequence effects is sometimes called item fungibility (Wainer & Mislevy, 2000) and is a prerequisite for any kind of method that flexibly combines calibrated items of an item pool. Item fungibility is often considered as an assumption of IRT (Thissen & Mislevy, 2000; Wainer & Mislevy, 2000; Wainer et al., 2007). However, this is misleading because IRT models with differing parameters for an item might be valid for the respective test forms. For instance, if sequence effects can be accounted for by a linear logistic test model (LLTM) with elementary position parameters, then the Rasch-model holds for each test form, respectively, but in general, difficulty parameters of an item do not concur for different test forms (Kubinger, 2009). This clearly demonstrates that sequence effects do not necessarily lead to violations of IRT, in general, or local stochastic independence, in particular.

Admittedly, context effects might not only change the parameters of the item but can also spoil the validity of the respective IRT model. For instance, all sequence effects that were introduced in the last paragraphs can result in local stochastic dependencies between item scores (Frey et al., 2009), i.e., the conditional probability of an item score can depend on the response that the test taker has given to another item. Local stochastic independence is an assumption for most models of item response theory (see van der Linden & Glas, 2010, for methods to test for local stochastic independence). However, models with locally dependent responses have been developed within the framework of IRT (e.g., Jannarone, 1997; Bradlow, Wainer & Wang, 1999; Tuerlinckx & De Boeck, 2004). Moreover, all kinds of sequence effects that were introduced in the previous paragraphs do not necessarily lead to local stochastic dependencies.

Furthermore, the consistency of item parameters across test forms does not guarantee that the same trait is measured by all test forms. Sequence effects (and context effects, in general) can be limited to a shift of the construct that is measured by the test forms. Theoretically, it is possible that the traits that are measured by different test forms with consistent item parameters do not even correlate. This point emphasizes the importance of item fungibility for the validity of educational and psychological tests that rely on calibrated item pools with item parameters that are not modeled as context-sensitive.

This line of argument implies that a scientifically sound development of a calibrated item pool requires getting evidence that sequence effects are negligible. In contrast, item fungibility is often simply postulated without any empirical evidence (Wainer et al., 2007). Admittedly, it is impossible to prove that there are no sequence effects at all. It is easy to imagine worst-case scenarios that preclude that sequence effects are revealed by any design for a study of sequence effects on item parameters. Nevertheless, it is possible to develop designs that can reveal strong evidence for the existence of sequence effects. Research in this field has shown that sequence effects are indeed an important issue (see Wainer et al., 2007, for an overview). If no evidence for sequence effects can be found in data on a specific item pool, then the assumption of item fungibility might be justified. Nevertheless, the complete absence of sequence effects will always be a postulate that is at best plausible. In fact, it would be naive to expect that there are no sequence effects at all even if they were not found in empirical data. The key question is if the size of these effects leads to serious threats to the validity of the test scores (see Schweizer, Reiss, Schreiner, & Altmeyer, 2012, for an example of validity improvement by eliminating position effects). Consequently, evidence for minor sequence effects does not necessarily thwart the development of a calibrated item pool as a base for flexibly composed test forms. However, the detrimental impact of sequence effects on the validity of test scores should be assessed and minimized.

Calibration designs

Goals (dependent variables) of a calibration design

Another issue is if and how minor sequence effects should be dealt with in the design and analysis of a calibration study. A *calibration design* describes in which way the items are assigned to the test forms and the test forms are subsequently assigned to test takers. A complete description not only refers to the question which items appear in which test form and which test forms are applied to which subpopulation of test takers but to the order of the items within a test form, too. If more than one test form is designated to a subpopulation, then the sampling scheme of test forms must be described, too.

Frey et al (2009) pointed out that the ultimate goal of a calibration study is to get unbiased and efficient estimates of item parameters. However, if there are sequence effects, it is not clear what is meant by an unbiased estimate of an item parameter because there are different parameters for that item in different test forms. In contrast, the notion 'unbiased' implies that there is one and only one genuine item parameter that is to be estimated. It has some appeal to define the item parameter as genuine that would lead to correct statements about the probability of an item score (for a test taker with known person parameter) if the response had been given without any impact of other items[9]. If item review is allowed, then the only test form for which the actual item parameter cannot be influenced by other items

[9] This concept of a genuine item parameter might be useful if the focus of test administration is on psychological interpretations of the item parameters.

is the test form that contains only this single item. However, it is hardly possible to get any sensible estimate for that item parameter because statistical techniques of item parameter estimation exploit the statistical association between item scores. If item review is not allowed, it is reasonable to assume that the item parameters are the same for all test forms in which this item appears as the first one because the following items cannot have any influence on the probability distribution of the item score.

Alternatively, the genuine parameter of an item can be defined as the mean of all parameters of an item across all possible test forms in which it is contained. Of course, this requires that the item parameters of all test forms can be brought on the same scale with a method that does not rely on equality constraints on item parameters. Efforts to balance sequence effects in the calibration study of an item pool can be understood as an attempt to ensure that the estimated item parameter is indeed the genuine item parameter defined as the mean across all possible test forms[10].

If a genuine item parameter is defined, then the bias of an item parameter within a calibration design can be defined as the expected value of the difference between the expected value of the item parameter estimate and the respective genuine item parameter. However, sequence effects are only one reason for a bias. Statistical estimation methods can cause biased item parameters, too. Simulation studies can reveal if a substantial bias can be expected due to statistical reasons.

Efficiency is the second goal of a calibration study that Frey et al. (2009) mentioned. In general, efficiency refers to the relationship between utility and costs. In the context of calibration designs utility can be considered as the precision of the item parameter estimates whereas the costs are related to the amount of data that has to be collected. If the amount of data is the same for different calibration designs then efficiency only depends on precision. The precision of a calibration design can be evaluated by considering the standard errors of the item parameters. Overall, indices of precision of a calibration design like the A-, D-, or E-optimality criterion can be derived from the variance-covariance matrix of item parameter estimates (Berger & van der Linden, 1991). These optimality criteria might be applied to all item parameters at once or to subsets of the item parameters separately, e.g. it can be more informative to look at these criteria for the discrimination and the difficulty parameters separately.

Principles and properties of calibration designs

So far we have only dealt with the outcome variables of calibration designs. In this section we want to give an overview over calibration designs. It would be hardly possible to enumerate and discuss all possible alternatives of calibration designs in one paper. Instead, we will outline general strategies and principles that underlie the development of a calibration design.

[10] However, techniques that include data for different test forms to derive a single item parameter estimate do usually not aggregate data in a way that ensures that this estimate is indeed modeled as mean item parameter across test forms.

The numerous alternatives of calibration designs trace back to two basic strategies that underlie the basic equating designs discussed in the literature, too (Vale, 1986; Dorans, Moses, & Eignor, 2011). These strategies ensure an adequate linkage of test forms.

1. The first strategy is to administer different test forms to equivalent groups of test takers. This allows exploiting the equality of person parameter distributions. This strategy relies on the postulate that the same trait is measured by all test forms. The equivalence of the groups of test takers can be ensured by randomization (test forms are assigned randomly to test takers) or by spiraling (test forms are alternated from one test administration to the next in a systematic manner). Spiraling guarantees equivalence of groups if test takers are sampled randomly from the respective population. In comparison to random sampling, combining spiraling with haphazard sampling of test takers can spoil or boost the equivalence of the groups. Spiraling of test forms with a balanced item sequence ensures that the respective aspects of the item sequence are counterbalanced across test takers, too. In any case, it should be taken care that the sampling of test takers is compatible with the respective (implicit or explicit) assumptions of the statistical model for item parameter estimation.
2. The second strategy is to assemble test forms with common items, which allows exploiting equality constraints on the respective item parameters across test forms. This strategy relies on item fungibility and measurement invariance (across the groups that receive different test forms) and can be thwarted by sequence effects and a lack of measurement invariance (e.g. differential item functioning). The set of common items is often called *anchor*. An anchor might also be called *anchor testlet* if the respective items are administered at a stretch in the same order in different test forms. A *global anchor* is part of every test form of a calibration design. A *local anchor* is part of a proper subset of test forms. Two test forms are directly linked if they share an anchor. If all test forms are directly or indirectly linked via anchors and item fungibility and measurement invariance hold, then the estimation of item parameters does not require that the test forms are administered to equivalent groups. This allows developing an ordered chain of linked test forms that measure one trait on a common scale. This is extremely useful if the overall population of test takers can or must be split into ordered subpopulations. This situation arises in vertical scaling if items are not suited for all subpopulations because of substantial differences in ability. These differences cause extreme score probabilities that thwart efficient item parameter estimation. Another example is a test program with cohorts of test takers who receive test forms that are linked over periods.

Calibration designs are more informative if they combine both strategies. If test forms with overlapping items are administered to equivalent groups (with a reasonable sample size), equality constraints on item parameters across test forms need not to be postulated

but become a testable assumption[11] (von Davier & von Davier, 2007, 2011). This line of argument can even be pursued if an ordered chain of test forms for ordered subpopulations is developed. If only one test form is administered to each subpopulation, then test form and subpopulation are confounded. Confounding can be attenuated by assigning one of two neighbored test forms (of the chain) randomly to each test taker of each subpopulation in a way that each test form is applied to members of two neighbored subpopulations. Then measurement invariance can be tested by comparing item parameters of a test form across neighbored subpopulations and sequence effects can be tested by comparing neighbored test forms that are applied to the same subpopulation.

Anchor designs apply various techniques to control for sequence effects on item parameter estimates with methods that mimic procedures of experimental design (Giesbrecht & Gumpertz, 2004):

- *Hold effect constant across test forms*
  Factors that might have an influence on the item parameters and the trait that is measured can be held constant across test forms. Hence, differences of the item parameters cannot be attributed to these factors. The drawback of this control technique is that generalizations to constellations in which these factors differ from the values that were held constant in the calibration study are not warranted. Therefore, factors that are held constant in the calibration design should be held constant in the operational stage of the test program, too. Calibration designs often combine this technique with other control techniques. For instance, all calibration designs that split the item pool into testlets hold sequence effects constant between items that appear in the same testlet. Moreover, in the remainder, we assume that other factors regarding the context (like the introductory part of the test forms and the mode of presentation) are held constant across all test forms.
- *Balance effects across test forms*
  Sequence effects are said to be counterbalanced if all values of the respective factor occur with the same frequency across all test forms. For example, the position of an item is balanced if this item appears on each position with the same frequency across all test forms. A calibration design is said to be first-order carry-over balanced if the following condition holds for all item pairs: The frequency of test forms with the first item of the pair as immediate predecessor of the second item matches the frequency of test forms with the first item as immediate successor of the second item of the pair. Another aspect that can be balanced across test forms is the common occurrence of item pairs across test forms regardless of the order and position in which they appear. This feature does not only allow for a homogenous linkage of test forms but also for an estimation of the statistical association between all items that enable analyses of model fit and of the dimensionality of the item pool. It should be kept in mind that balancing an aspect of the item sequence across test forms of a calibration design does not necessarily guarantee that the effects of the

[11] Interestingly, the only sequence effects that largely elude empirical testing refer to shifts in the construct that is measured. Fortunately, it is not very plausible to assume that shifts in the construct due to sequence effects leave the item parameters unchanged.

respective factor on the estimation of item parameters are counterbalanced. This would require a technique of aggregating data from different test forms in a way that ensures that the item parameter estimate actually refers to the specified genuine item parameter. Nevertheless, we loosely speak of counterbalanced sequence effects throughout this paper when the respective aspect of the sequence is balanced.

- *Randomize effects across test forms*
  Balancing requires that the values of each factor or each combination of factors that could have an influence on the estimated item parameters are realized at least once. However, the number of possible combinations of factor values that could have an influence on the item parameters rapidly becomes prohibitively large. A common technique of experimental design and statistics to overcome this problem is random sampling. Random sampling can be applied at various stages of a calibration study. It can be applied to draw test takers from an infinite (sub-)population of (potential) test takers. It can be applied to assign items (or testlets) to test forms and positions within a test form, and, finally, it can be applied to assign test forms to test takers. The latter procedure mimics randomization of conditions in experimental design. In any case, random sampling cannot ensure that a factor is counterbalanced exactly. However, random sampling ensures that the respective factor is counterbalanced approximately, i.e., with increasing sample size perfect counterbalancing is approached. Statistical methods can be developed to account for random confounding that might still be present.
- *Statistical control of sequence effects across test forms*
  Another method to control for sequence effects is to incorporate these effects in the statistical model that is used to analyze the data. Within such an approach the parameter of an item in a test form would be decomposed into the genuine item parameter and parameters that take account for the sequence effect (Kubinger, 2008, 2009).
- *Blocking*
  In the broadest sense, blocking refers to any procedure that splits a finite or infinite set (or population) into mutually exclusive and exhaustive subsets (or subpopulations). Blocking is a common technique in experimental design that groups experimental units sharing a common feature. It potentially reduces irrelevant variation and allows for a more precise estimation of effects. Blocking of the population (or the sample) of test takers is used in multiple-group models of IRT to ensure unbiased estimation of IRT parameters. In the field of educational measurement it often happens that a blocking factor (like a cohort of students) is completely confounded with the test form that is administered. A common feature of calibration designs is the blocking of items by dividing the item pool into non-overlapping testlets. In fact, all calibration designs that will be discussed in the remainder can be applied either to items or to non-overlapping testlets. Hence, the term *unit* will be used to address both options. Counterbalancing the sequence of all testlets instead of balancing the sequence of all items substantially reduces the number of test forms of a calibration design which boosts practicability; especially if the test is presented in paper-and-pencil format. Decreasing the number of test forms yields a bigger sample size for each test form which boosts the power of statistical tests of equality constraints on

item parameters across test forms. In fact, if the number of test forms is not substantially lower than the overall sample size of test takers equality constraints must be postulated without any empirical test. Moreover, sequence effects are held constant within the testlet which can attenuate the problem of sequence effects if the respective testlets are used as building blocks for the operational test forms, too. However, calibration designs with a balanced sequence of testlets usually do not counterbalance the sequence of items. Consequently, they do not offer an adequate control of sequence effects if items are used as elements for composing test forms in the operational stage of the test program.

Categories of calibration designs

Before we introduce concrete anchor calibration designs we like to stress two common features that are shared by all of the designs that will be discussed in this section:

- Each test form consists of the same number of units (i.e., items or testlets).
- Each unit occurs at most once in each test form.

The most simple calibration design is to develop just a *single test form* that is composed of a fixed sequence of all units (i.e., items or testlets) in the item pool. This calibration design is the standard for the development of a fixed test. In this case the main advantage is that all sequence effects that might have an influence on the item parameters are held constant not only within the calibration design but also across the calibration study and the operational stage of the test program. However, this feature of the most basic calibration design turns out to be a great disadvantage when applied to the development of a calibrated item pool that is intended as a basis for a flexible compilation of test forms. Item content and item context effects with respect to the sequence of the items are confounded. This confounding needs not to be considered as a serious problem if it is exactly replicated in the operational stage of the test program like in the development of a fixed test form. (Nevertheless, sequence effects can be considered as a possible threat to validity of the test because they might affect the adequacy of interpretations of item parameters and test scores.) However, if the sequence of the items of an operational test form does not match the sequence of the items of the test forms that are used for the calibration study, sequence effects on item parameters are an important issue that should always be addressed in the development of the item pool.

Calibration designs are said to be complete if every test taker responds to every item of the item pool. Consequently, only the order of items can vary across test forms. The single test form design that was introduced in the last paragraph is an example of a complete calibration design. A *complete item/testlet permutation design* is a calibration design that realizes all possible complete sequences of units (i.e., items or non-overlapping testlets) of the item pool, which ensures that all aspects of the sequence of units are counterbalanced. This is only feasible if the item pool consists of a limited number of units because the number of test forms would be $n!$ (*n factorial*) if the item pool contained n units. In contrast, the num-

ber of test forms of a *latin-square item/testlet calibration design* equals only the number of units (cf. Table 1a). Latin-square designs control for position effects, whereas balanced latin-square calibration designs (cf. Table 1b) control for first-order carry-over effects, too. Complete calibration designs are only suited for small item pools where test length (and consequently the burden for each test taker) remain at a reasonable level.

Incomplete calibration designs administer only a part of the item pool to each test taker. Responses to the remaining items are missing by design and hence completely at random which ensures that the statistical methods used to estimate the item parameters work properly (Mislevy & Wu, 1996). The number of units k of the test form is usually held constant across all test forms of an incomplete calibration design.

An *incomplete item/testlet permutation design* is a design that realizes all possible sequences of $k < n$ units of the item pool, which ensures that all item sequence effects are balanced. Although the number of possible test forms $n! / (n-k)!$ is usually considerably smaller than for complete permutation designs, it is still prohibitively large unless the number of units is very small. Nevertheless, the incomplete permutation is an important benchmark for other incomplete calibration designs with respect to balancing the unit sequence.

Table 1a:
Latin-Square Design

	test form			
position	**A**	**B**	**C**	**D**
1st	1	2	3	4
2nd	2	3	4	1
3rd	3	4	1	2
4th	4	1	2	3

Note: The numbers correspond to the units (testlets or items).

Table 1b:
Balanced Latin-Square Design

	test form			
position	**A**	**B**	**C**	**D**
1st	1	2	3	4
2nd	2	3	4	1
3rd	4	1	2	3
4th	3	4	1	2

Note: The numbers correspond to the units (testlets or items).

The *random item/testlet design* usually reaches a comparable level of balancing by administering a random sample (without replacement) of units of the item pool to each test taker. All sequence effects are approximately balanced. Unless the number of units is very small, most possible test forms will not be administered at all and the remaining test forms will usually only be administered to a single test taker. The common occurrence of each pair of units is counterbalanced. If it is intended to use item correlations for analyzing model fit and dimensionality of the item pool, then the sample size of test takers should be sufficiently large so that enough data for each pair of units is available. Simulation studies should precede data collection to check if problems due to the sparseness of the data matrix should be expected.

An *interlaced item/testlet design* realizes all coherent subsets (of size *k*) of a cyclic order of the units of the item pool (cf. Table 1c). Consequently, the number of test forms equals the number of units. The position of the units is counterbalanced. The strength of the interlaced *item* design is that first-order carry-over effects are nearly held constant and a strong linkage of "neighbored" test forms is established. The interlaced *testlet* design is often implemented with only two testlets per test form (cf. Table 1d). This results in a relatively small number of test forms for the given test length. The number of test forms can only be reduced further by lowering the linkage (item overlap) across test forms. The most economic calibration design with respect to the necessary test forms for a fixed value of test length is the *unlinked test forms design* (cf. Table 1e) that does not rely on anchor items but completely on the equivalence of groups that receive the different test forms.

A *balanced incomplete block design* (BIBD) ensures that across all test forms the concurrence of each pair of units is the same (cf. Table 1f). BIBDs do not necessarily balance position and carry-over effects. A Youden square is a special position balanced BIBD that ensures that each unit appears once at each position, which always results in a number of test forms that equals the number of units (cf. Table 1g). In general, Youden squares are not carry-over balanced. BIBDs that are position and first order carry-over balanced can be generated by the package *crossdes* (Sailer, 2005) of the open-source statistical software R. Table 1h gives an example of a position and carry-over balanced design that is no BIBD where each unit appears twice at each position. A common feature of all calibration designs that were introduced in this paragraph is that they exist only for certain combinations of test length and of the overall number of units in the item pool (Cochran & Cox, 1957; Giesbrecht & Gumpertz, 2004; Frey et al., 2009).

Table 1c:
Interlaced Design

	test form				
position	**A**	**B**	**C**	**D**	**E**
1st	1	2	3	4	5
2nd	2	3	4	5	1
3rd	3	4	5	1	2

Note: The numbers correspond to the units (testlets or items).

Table 1d:
Interlaced Testlet Design with Minimal Linkage

	test form				
position	**A**	**B**	**C**	**D**	**E**
1st	1	2	3	4	5
2nd	2	3	4	5	1

Note: The numbers correspond to the units (testlets or items).

Table 1e:
Design with unlinked Test Forms

	test form		
position	**A**	**B**	**C**
1st	1	4	7
2nd	2	5	8
3rd	3	6	9

Note: The numbers correspond to the units (testlets or items).

Table 1f:
Balanced Incomplete Block Design (BIBD)

	test form					
position	**A**	**B**	**C**	**D**	**E**	**F**
1st	1	1	1	2	2	3
2nd	2	3	4	3	4	4

Note: The numbers correspond to the units (testlets or items).

Table 1g:
Youden Square Design

	test form			
position	**A**	**B**	**C**	**D**
1st	1	2	3	4
2nd	4	1	2	3
3rd	3	4	1	2

Note: The numbers correspond to the units (testlets or items).

Table 1h:
Position and (First-Order and Second-Order) Carry-Over Balanced Design

	test form									
position	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
1st	1	2	3	4	5	3	4	5	1	2
2nd	5	1	2	3	4	5	1	2	3	4
3rd	3	4	5	1	2	1	2	3	4	5

Note: This is no balanced incomplete block design. The numbers correspond to the units (testlets or items).

Table 2 gives an overview over properties of calibration designs. Of course this list of calibration designs is not complete and combinations of these design types are possible, too. The decision for a calibration design has to take into account the goals of the study and practical constraints. A general recommendation is to maximize the similarity of context factors of the calibration study and the operational stage of a test program.

Table 2:
Counterbalancing and Completeness in various Calibration Designs

complete	**counterbalanced**					**design**
	position	**carry-over**		**concurrence**	**occurrence**	
		first-order	**higher-order**			
yes	–	–	–	++	++	one fixed test form
yes	++	++	++	++	++	complete permutation
yes	++	–	–	++	++	latin-square
yes	++	++	0	++	++	balanced latin-square
no	++	++	++	++	++	incomplete permutation
0	+	+	+	+	+	random
0	++	–	–	0	++	interlaced
no	0	0	0	++	++	balanced incomplete block design
no	++	–	–	++	++	Youden square

Note: ++: counterbalanced exactly; +: counterbalanced approximately; –: not counterbalanced; 0: not specified.

Empirical example

Frey et al. (2009) claim a lack of theory for calibration designs. The preceding chapters can be understood as an attempt to provide a nomenclature for calibration designs with special attention to the problem of sequence effects. The following example demonstrates how this nomenclature can guide the process of developing a calibrated item pool in an applied setting.

The purpose of the drafted project is the development of a number series test as a measure for numerical reasoning. The test is intended as a tool for psychological expertise in the Psychological Service of the German Federal Employment Agency. Test takers are adolescents and adults and the whole range of numerical reasoning ability is targeted. The population of test takers is divided into four subpopulations that refer to the respective educational attainment[12]. Because of the wide range of ability of the total population of test takers, it is necessary to adapt the difficulty of the administered items to the ability of the respective test taker to avoid detrimental impacts on the reliability and validity of test scores. To date this is done by administering different fixed tests which are not linked onto a common scale. The goal of the project is to develop a computerized adaptive test that is suited for the whole range of ability. Items should be selected from a final item pool of 100 to 120 items in order to achieve adequate measurement accuracy and a basic level of item security. Expectations with regard to the distributions of item and person parameters were derived from experience with other CAT item pools that are already implemented.

Rules for item generation were specified after inspection of published number series tests and the mathematical and psychological literature on number series (e.g., Holzman, Pellegrino, & Glaser, 1983; Korossy, 1998). An item generator was developed that is able to generate a class of number series items by specifying some basic parameters. The resulting number series items were automatically evaluated with respect to several criteria (so-called radicals; Irvine, 2002) that were intended to facilitate an expert ranking of item difficulty. These expert rankings were based on radicals derived from relevant literature on number series items (e.g. Holzman et al., 1983) and from the analysis of available empirical data of other number series tests. The approximate sample size of test takers is determined by organizational constraints. Two simulation studies were run before the final calibration design was determined. The main purpose of the simulation studies was to figure out a design that controls for sequence effects and allows assessing the size and nature of sequence effects. Nevertheless, the design should be optimal with respect to item parameter recovery in case of item fungibility, which is expected to hold approximately for the item pool. An IRT model with item difficulty and discrimination parameters (2-pl) was used to analyze the data[13].

---

[12] Adolescents are grouped by the expected educational attainment if they still attend school.

[13] The 2-pl model was chosen because guessing should not be an issue with number series items with open response format. It should be kept in mind that the results of the simulation studies might have been different if a 1-pl or a 3-pl model had been used.

Simulation study 1

Vale (1986) reported that the accuracy and the bias of the item parameter estimates do not depend on the choice of the anchor design if test forms are applied to equivalent groups. The goal of the first simulation study was to replicate this result for the equivalent groups calibration designs that were initially considered for the final calibration study. Moreover, the bias and the accuracy of the item parameter estimates should be analyzed as a function of the true item parameter values in order to see if an acceptable level of accuracy can be reached with the planned sample size of test takers. It was expected that the item parameter estimates are unbiased due to the large sample size. Moreover, the standard error of item parameters can be expected to increase with the deviation of the item difficulty parameter from the mean of the person parameter distribution (Berger & van der Linden, 1991).

Method. Within each design all test forms consisted of 15 items. The following anchor designs were compared (cf. Table 3):

- an unlinked test form design,
- an interlaced testlet design with local anchors of 7 and 8 items, respectively,
- an interlaced testlet design with local anchors of 5 items and an additional global anchor of 5 items,
- a balanced incomplete block design that is counterbalanced with respect to first-order carry-over and position effects with testlets of 5 items,
- a random item design,
- and a single test form design.

The overall amount of data, i.e., the number of responses across items and test takers was the same for all anchor designs.

Properties of the calibration designs are summarized in Table 3. Note that the single test form and the unlinked test form design neither balance sequence effects nor do they allow to model or analyze them. Each item appears in a limited number of test forms in the balanced incomplete block design and both interlaced testlet designs. If the parameter estimates of an item vary across test forms, then there is evidence for sequence effects. In contrast, there are at best only a few observations (i.e., test takers) for each test form that is realized in the random item and the random testlet design. Consequently, it is not possible to estimate test-form specific item parameters. Nevertheless, evidence for sequence effects might be derived by adjusting an IRT model that incorporates parameters for sequence effects.

Table 3:
Design Features in Simulation Study 1 with Results

	design features									results (averaged across items and conditions)			
	test length		reponses/item		counterbalanced					bias		RMSE	
design	Min	Max	Min	Max	position	carry-over: first-order	carry-over: higher-order	concurrence	can detect sequence effects	*a*	*b*	*a*	*b*
complete	125	125	3600	3600	–	–	–	–	No	-.012	.001	.069	.074
BIBD position and carry-over balanced	15	15	3600	3600	+	+	–	+	Yes	-.007	.000	.082	.077
random item	15	15	3487	3745	++	++	++	++	Limited	-.008	.000	.082	.087
random testlet	15	15	3510	3716	+	+	+	+	Limited	-.008	.001	.083	.082
interlaced testlet	15	16	3600	3600	+	–	–	–	Yes	-.007	.000	.082	.079
interlaced testlet with global anchor	15	15	2500	30000	–	–	–	–	Yes	-.001	-.000	.096	.127
unlinked	15	16	3600	3600	–	–	–	–	No	-.007	-.001	.083	.082

Note: ++: counterbalanced for items; +: counterbalanced for testlets; –: not counterbalanced; RMSE: root mean squared error; a: item discrimination parameter; b: item difficulty parameter

Item difficulty parameters (b) were drawn from a standard normal distribution, item discrimination parameters (a) from a log-normal distribution ($log(a) \sim N(0.25, 0.25)$). Both item parameters for the 125 items were only drawn once and held constant over the runs of the Monte-Carlo simulation. The distribution of the item parameters is shown in Figure 1. Person parameters were drawn from a normal distribution. Three different values of the mean (-1, 0, 1) and three different values of the standard deviation (0.75, 1, 1.4) of the respective distribution were realized in order to see if the match to the distribution of item difficulty matters. Varying the distribution parameters of the person parameter distribution is equivalent to variations of the scale of the item parameters while the pattern of item parameters is held constant. 15 replications were performed for each combination of factors. Item parameters were estimated with the marginal maximum likelihood method, i.e., no prior distribution was specified for the item parameters. Data was simulated by the R software and estimation was done with BILOG-MG with 20 quadrature points.

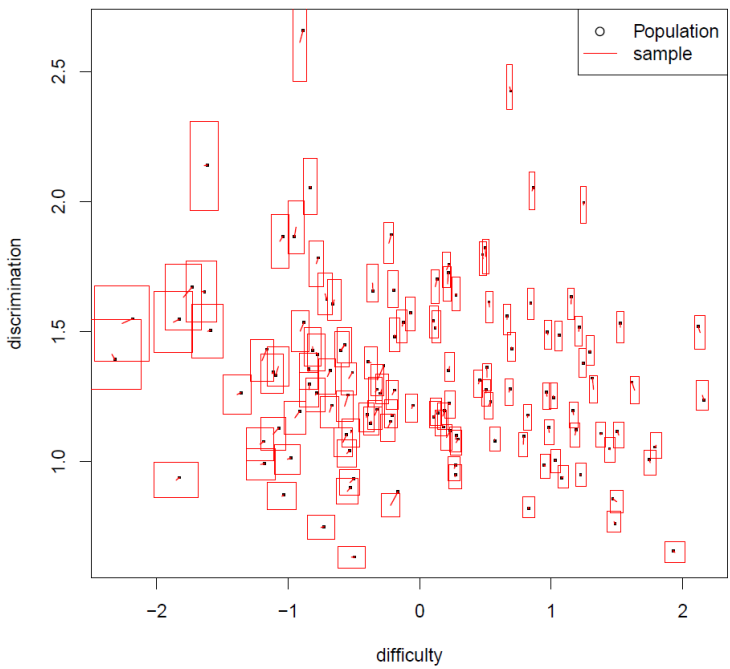

Figure 1:
Bias and standard errors of item parameter estimates for the design "complete" with a person parameter distribution of N(1,1). The dots correspond to the population values of the item difficulty and discrimination parameters. They are connected with the mean of the item parameter estimates that is in the center of a rectangle while the distance from the center to the edges corresponds to the standard error.

Results. Item difficulty was estimated with a slight outward bias, i.e., the correlation of item difficulty with the bias of item difficulty estimates was about .50 (cf. Table 4). However, the size of the bias of item difficulty and item discrimination was negligible (≤ .02 on average across all items). The accuracy of the item parameter estimates was slightly better for the complete design (cf. Table 3). There were only minor differences with regard to the accuracy between the incomplete calibration designs (a typical pattern is depicted in Figure 1). On average, across all items the lowest accuracy of item parameter estimates was observed for the interlaced testlet design with an additional global anchor. However, the accuracy of the global anchor items was better in comparison to the other designs. This observation could be attributed to the fact that the amount of data is not evenly spread across all items for this calibration design. In general, anchor items did not lead to an increase of the stability of the item parameter estimates. The accuracy of the item discrimination estimates was negatively related with the true item discrimination parameter whereas the relation with the precision of the item difficulty estimates was positive. The most striking result was that the standard error of the item parameters of items with a (true) difficulty parameter that lies at the extremes of the person parameter distribution was larger (cf. Figure 1).

Conclusions. The results of the simulation study are in line with the results of Vale (1986) who reported that the anchor design is irrelevant if test forms are applied to equivalent groups. Therefore, it was decided to build on the random item design because it offers the best control of potential sequence effects. The data generation model of the simulation study assumed item fungibility. Nevertheless, it seems reasonable to control for potential sequence effects. However, item parameter estimates of items with extreme difficulty were not satisfactory for all equivalent groups calibration designs which can be attributed to extreme score probabilities for most test takers. Consequently, principles of

Table 4:
Correlation of the Bias of the Item Difficulty Estimates with the true Item Difficulty

distribution parameters of the person parameter distribution		***r***
M	***SD***	
-1	.75	.39
	1	.52
	1.4	.52
0	.75	.51
	1	.49
	1.4	.53
1	.75	.46
	1	.47
	1.4	.48

vertical scaling potentially ameliorate item parameter estimates by administering only those parts of the item pool that are tailored to the level of ability of the respective subpopulation of test takers. Therefore, a second simulation study was run in order to analyze if parameter estimates of items with extreme difficulty can be improved by vertical scaling designs.

Simulation study 2

The item difficulty ranking was used to split the item pool into overlapping clusters ordered with respect to item difficulty. This allows for a calibration design that administers items of adequate difficulty to the respective subpopulation of test takers. Thereby, a key feature of the CAT, namely the adaption of the difficulty level of the items to the level of ability, could be introduced into the calibration design. As pointed out before, maximizing the similarity of context factors of the calibration design and the operational test is a major strategy to minimize the bias of item parameter estimates of the operational test form that is eventually administered.

Vertical scaling designs administer an ordered chain of test forms to subpopulations that are ordered with respect to ability. However, the test forms would become prohibitively long if all items of the subpopulation-specific item pools were administered together. It is straightforward to administer only a random sample of the respective item pool. Actually, the designs that were compared in this simulation study administer a random item design to each subpopulation with a subpopulation-specific item pool. Each subpopulation-specific item pool corresponds to an interval of ranks of the (overall) item pool with respect to item difficulty. The intervals of ranks of neighbored subpopulation-specific item pools are overlapping, i.e., the item pools are linked. Hence, the designs compared in this simulation study are combinations of random item and vertical scaling designs, respectively. They (potentially) prevent the administration of items that are inappropriate for the respective subpopulation of test takers due to an extreme probability distribution of the item score. These items offer only negligible information for the estimation of item parameters and potentially induce frustration and demotivation which might spoil the validity of test scores.

The designs compared in this study differ with respect to the degree of attunement of item difficulty to ability. From a substantive point of view the degree of attunement should be maximized in order to maximize similarity to the operational CAT and to minimize the risk of administering items that are not suited for the respective test takers. The main purpose of the simulation study was to compare the psychometric properties of the respective calibration designs.

Method. Person parameters were sampled from normal distributions. The subpopulation-specific distribution parameters and the size of the samples of test takers corresponded to values from previous empirical studies within the organization (cf. Table 5).

Table 5:
Size and Distribution of Ability within Subpopulations/Subsamples

educational level	*N*	*M*	*SD*
1	8000	-1.10	0.80
2	18000	-0.18	0.88
3	9000	0.37	0.92
4	2000	1.47	1.02

Difficulty (b) and discrimination (a) parameters of 150 items were sampled from a normal and a log-normal distribution, respectively ($b \sim N(0,1)$, $log(a) \sim N(0.20,0.25)$). A rank order of the items with respect to difficulty was established. Subpopulation-specific item pools were composed of items that correspond to an interval of ranks, respectively. The size of the subpopulation-specific item pools is inversely related to the expected number of test takers in order to ensure that there is enough data for the estimation of all item parameters. The linkage of the items pools is depicted in Figure 2. The calibration designs differ with respect to three (confounded) factors:

- the amount of data (i.e., responses) that is cumulated for items with extreme difficulty,
- the fit of item difficulty to the ability of test takers,
- and the linkage (number of anchor items) between the subpopulation-specific item pools.

The design "weak" has the weakest linkage of item pools and the best attunement with respect to ability. The linkage is somewhat stronger in the design "strong" because every item except for the most extreme ones are included in an anchor. The design "ends" uses the most extreme items as anchors, too. This reduces the attunement to ability further. The design "three" differs from an equivalent groups random item design only by administering the most extreme items to the respective subpopulations with an extreme ability level. The design "single" (which is not depicted in Figure 2) is actually a random item design without any vertical scaling feature. It is included in the study for comparison in order to assess if the goals that motivated the application of vertical scaling were reached.

Results. The precision of the item parameter estimates of the different calibration designs is depicted in Figure 3. Descriptive statistics of the standard error of the item parameter estimates are reported in Table 6 together with the respective correlations with the true values of these parameters. As expected, the precision of parameter estimates of items with extreme difficulty was best for the vertical scaling designs that accumulated most data for these items (namely designs "three" and "ends"). The other vertical scaling calibration designs (i.e., "weak" and "strong") were not superior to the random item

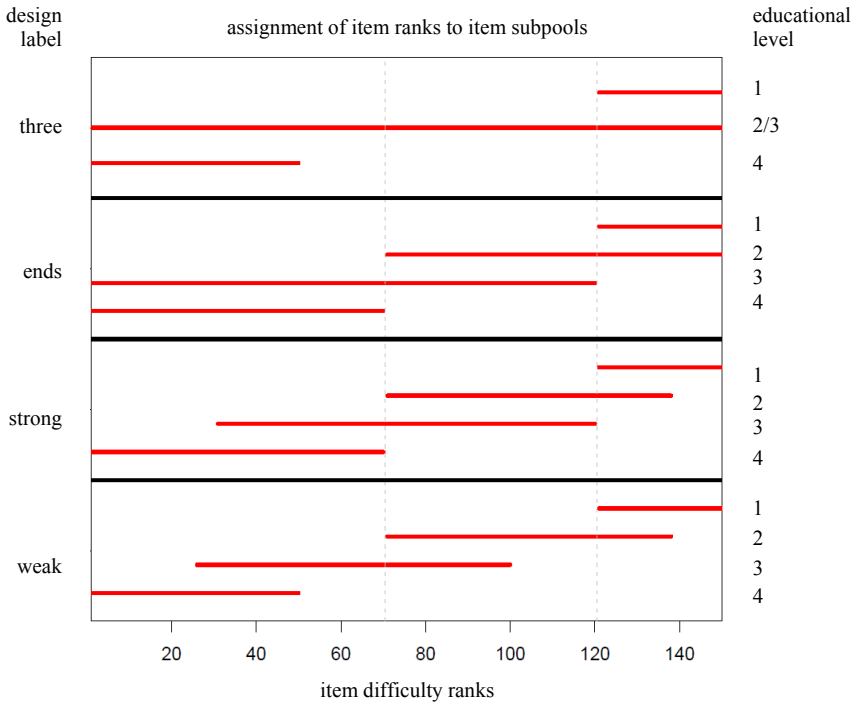

Figure 2:
Assignment of item ranks to subpopulation-specific item pools in simulation study 2. Educational level is numbered in descending order of the German school system: 1 = Gymnasium; 2 = Realschule; 3 = Hauptschule; 4 = Förderschule für Lernbehinderte. Note that each test taker does not respond to all items of the respective subpool but only to a random sample of 16 items.

design ("single") with respect to the accuracy of the parameter estimates of items with extreme difficulty. The vertical scaling design that led to the most heterogeneous pattern of accuracy was the design "weak". Accuracy was best for anchor items of medium difficulty and worst for items with extreme difficulty that appear only in one population-specific item pool. The design "ends" was most successful in limiting the standard error of item parameter estimates. The design "three" was most successful in limiting the correlation of the true absolute item difficulty with the standard error of the item parameter estimates.

Table 6:
Descriptive Statistics of the Distribution of the Standard Error and Correlations of the Standard Error with the respective true Values

stand-ard error	design	descriptive statistics of the distribution of the standard error							correlation of the standard error with true values		
		M	*SD*	Min	Q25	Median	Q75	Max	*a*	*b*	$\lvert b \rvert$
SE_a	single	.060	.015	.038	.049	.056	.066	.134	.86	.06	.31
	three	.061	.016	.033	.048	.060	.069	.123	.69	.17	-.38
	ends	.059	.012	.039	.050	.056	.065	.099	.89	.12	.21
	strong	.066	.025	.036	.047	.055	.085	.156	.41	-.15	.69
	weak	.066	.023	.032	.049	.060	.079	.164	.40	.22	.59
SE_b	single	.048	.023	.023	.031	.040	.058	.146	-.45	.29	.84
	three	.041	.014	.019	.030	.038	.049	.086	-.66	.36	.48
	ends	.038	.014	.021	.028	.034	.047	.097	-.62	.16	.71
	strong	.040	.019	.019	.027	.034	.047	.122	-.47	.00	.82
	weak	.038	.023	.017	.025	.033	.047	.104	-.54	.26	.74

Note: a = item discrimination parameter; b = item difficulty parameter; $\lvert b \rvert$ = absolute value of item difficulty parameter.

Conclusions. From a psychometric point of view the design "ends" (cf. Figure 2) is superior to the other designs because it does not only lead to a low level of standard errors but also to the most homogenous pattern of standard errors across the whole range of item difficulty and item discrimination. However, from a substantive point of view it must be noted that the designs "weak" and "strong" offer a better level of attunement of item difficulty to the ability of the test takers in the respective subpopulation. Like mentioned above, this feature minimizes threats to the validity of test scores due to effects like frustration and inattentiveness. These effects are not modeled in the simulation study, but they might be a problem when administering calibration designs to real samples of test takers. In summary, a universal recommendation cannot be derived but a compromise has to be found between psychometric accuracy and substantive appropriateness that should take into account the practical constraints of the respective test program.

Description of the final calibration design

The appropriateness of the items is of special concern for the calibration design of the number series test because the data is gathered in a high-stakes setting, but demotivation is nonetheless a common problem with the respective clientele, especially at the lowest educational level. Consequently, the design "strong" was chosen because it offers a

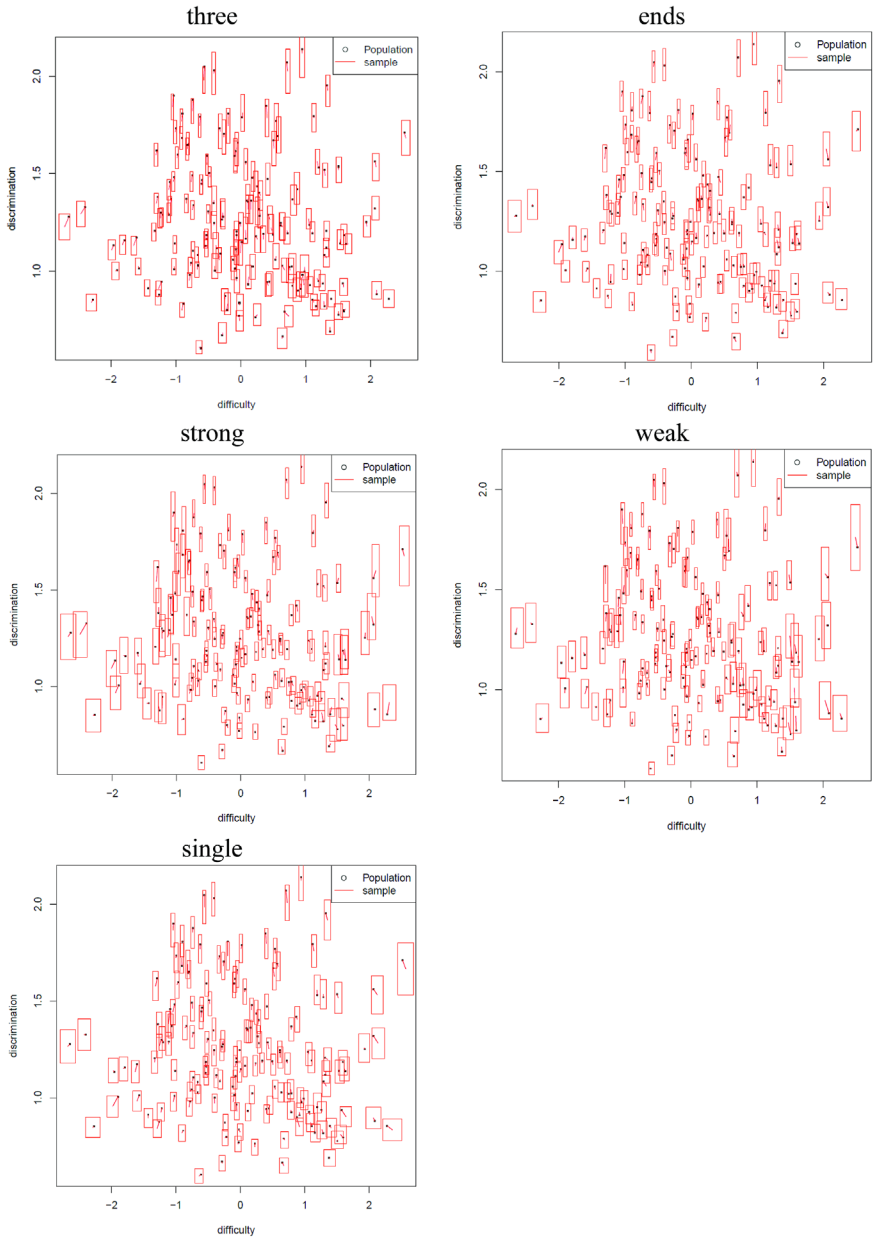

Figure 3:
Comparison of the item parameter estimates of the calibration designs in study 2. The dots correspond to the population values of the item difficulty and discrimination parameters. They are connected with the mean of the item parameter estimates that is in the center of a rectangle while the distance from the center to the edges corresponds to the standard error.

sufficient level of accuracy of the item parameter estimates and strictly avoids administering items which might not be appropriate for the respective subpopulation. The latter property is also a main feature of the planned operational CAT. Thereby, the similarity of context factors (here: the difficulty of the other items in the test forms) is maximized.

The design "strong" is a stratified random item design that counterbalances all kinds of sequence effects within the item subpools. Moreover, the resulting item parameter estimates can be considered as good proxies for the genuine item parameters defined as the means of item parameters of all test forms that can be composed of the respective item subpool. However, it is not easy to assess the size and the nature of potential sequence effects with this calibration design.

It is a hopeless endeavor to screen for all possible sequence effects that could arise when composing test forms of a large item pool. For this reason, it was decided to supplement the design "strong" with experimental test forms targeted at testing specific hypotheses about the nature of potential sequence effect by comparing the respective item parameters with the genuine item parameters resulting from the stratified random item design. Therefore, the calibration design is supplemented by fixed test forms that are administered to random subsamples of test takers. These test forms are implemented for experimental purposes and are not intended to be used for the calibration of the item pool. They follow certain principles that allow testing hypotheses about the nature and the relevance of sequence effects for test assembly:

- *Increasing difficulty*
  This principle guides the composition of most fixed test forms. If the item parameters of test forms that follow this principle differ substantially from the genuine item parameters of the respective random item design, then it is inappropriate to assemble operational tests from the calibrated item pool that follow this principle.
- *Homogenous difficulty*
  The difficulty of the presented items of computerized adaptive tests tends to stabilize quickly. If the item parameters of test forms with homogenous difficulty differ markedly from the parameters that were estimated by the stratified random item design, then it would seem questionable if the estimated item parameters of the calibrated item pool are actually applicable for the CAT presentation mode.
- *Fixed unsystematic order*
  The test form was realized as an attempt to get a hint about the variability of item parameters of test forms that are not ordered with respect to difficulty. Specific deviations to the genuine item parameters are not expected because this test form implements an unsystematic mode of item selection as well.
- *Reversed unsystematic order*
  This test form presents the same items as the test form with fixed unsystematic order but in reversed order for the purpose of analyzing position and carry-over effects.

The first two principles were based on the expert ratings of item difficulty that were already mentioned. The pattern of the differences of the item parameters across the ex-

perimental test forms and the genuine random item vertical scaling calibration design are expected to shed light on the nature and the size of potential sequence effects. If the item parameter estimates from the described calibration study show evidence for non-negligible sequence effects simulation studies will be run to estimate the practical consequences for the validity of the test scores.

Discussion

The issue of sequence effects has been acknowledged since test developers began to administer test forms that are composed of subsets of an item pool. These concerns accompanied the development of multi-matrix-sampling and computerized adaptive testing and were a major inspiration for the development of multistage testing. Moreover, there is evidence that sequence effects are not only possible in theory but are a phenomenon that can be found in real data (Wainer et al., 2007; Hohensinn et al., 2008). Nevertheless, test developers seldom address these concerns in an adequate way but tend to postulate item fungibility. Contrariwise, there seems to be a preference for calibration designs that counterbalance sequence effects. However, if the assumption of item fungibility holds, then it is a needless exertion to counterbalance sequence effects. If there are substantial sequence effects, then none of the resulting problems is remedied by counterbalancing the item sequence in the calibration study.

A scientifically sound interpretation of test scores that stem from test forms that are composed of a calibrated item pool requires that the respective calibration study implemented empirical tests of sequence effects. It might be argued that it is a hopeless endeavor to verify item fungibility for item pools that contain more than a few items. Nevertheless, sequence effects are amenable to empirical testing, i.e., a calibration study with an adequate design can exhibit strong evidence for sequence effects. However, the key question of the empirical analysis of sequence effects is not if they exist but if they are substantial. If there is evidence for substantial sequence effects, then all applications that rely on item fungibility will probably lead to invalid inferences. One way (if not the only) to remedy this threat to validity might be to include the substantial sequence effects in the measurement model.

If the data does not indicate substantial sequence effects, then the benefit of counterbalancing the item sequence narrows to minimizing small distortions due to minor sequence effects. However, it should be kept in mind that counterbalancing the item sequence does not necessarily lead to counterbalanced sequence effects. The expected value of an item parameter estimate in a random item design (with concurrent calibration of the random test forms) needs not to be the mean of item parameter values across all test forms (although it might usually be a good proxy). This line of argument highlights the necessity of the concept of a genuine item parameter as a benchmark for item calibration studies if item parameters depend on the test form. It depends on the concrete application how the genuine item parameter is defined best. The mean of the parameters of an item across all test forms appears to be a natural choice if all test forms are administered with the same probability. This needs not to be the case in every application. For instance, CAT tends

to compose test forms with homogenous item difficulty because the person parameter estimate usually stabilizes rapidly during the test administration. A general strategy to minimize detrimental impacts of context effects on the validity of test scores is to maximize the similarity of context factors (like aspects of the item sequence) between the calibration study and the operational stage of a test program.

Acknowledgement

We are grateful to Peter M. Muck, Steffi Pohl, and two anonymous reviewers for helpful comments on an earlier version of this manuscript.

References

Berger, M. P. F. & van der Linden, W. J. (1991). Optimality of sampling designs in item response theory models. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 274-288). Norwood, NJ: Ablex Publishing Company.

Bock, R. D. & Zimowski, M. F. (1997). Multi group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433-448). New York: Springer.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

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

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21-42). New York: Springer.

Eggen, T. J. H. M & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30,* 379-393.

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

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

Gonzalez, E. & Rutkowski, L. (2009). Principles of multi matrix booklet designs and parameter recovery in large-scale assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2*, 9-36.

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly, 50,* 391-402.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation, 17,* 497–509.

Holzman, T. G., Pellegrino, J. W., & Glaser, R. (1983). Cognitive variables in series completion. *Journal of Educational Psychology, 75,* 603-618.

Irvine S. H. (2002). The foundations of item generation for mass testing. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 3–34). Mahwah, NJ: Lawrence Erlbaum Associates.

Jannarone, R. J. (1997). Models for locally dependent responses: Conjunctive item response theory. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 465-480). New York: Springer.

Khorramdel, L. & Frebort, M. (2011). Context effects on test performance: What about test order? *European Journal of Psychological Assessment, 27,* 103-110.

Kolen , M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York: Springer.

Korossy, K. (1998). Solvability and uniqueness of linear-recursive number sequence tasks. *Methods of Psychological Research Online, 3,* 43-68.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From composing tests by item generating rules to measuring item administration effects. *Psychology Science Quarterly, 50,* 311-327.

Kubinger, K. D. (2009). Applications of the Linear Logistic Test Model in psychometric research. *Educational and Psychological Measurement, 69,* 232-244.

Le, L. T. (2009). 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* (pp. 217-226). Tokyo: Universal Academic.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Mislevy, R. J. & Wu, P. K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, adaptive testing.* ETS Research Report RR-96-30-ONR. Princeton, NJ: Educational Testing Service.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York: Springer.

Rost, J. (1997). Logistic mixture models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449-464). New York: Springer.

Sailer. O. (2005). crossdes: A package for design and randomization in crossover studies. *R News, 5(2)*, 24-27.

Schweizer, K., Reiss, S., Schreiner, M., & Altmeyer, M. (2012). Validity improvement in two reasoning measures due to the elimination of the position effect. *Journal of Individual Differences, 33*, 54-61.

Thissen, D. & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101-133). Mahwah, NJ: Lawrence Erlbaum Associates.

Tuerlinckx, F. & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Eds.), *Explanatory item reponse models. A generalized linear and nonlinear approach* (pp. 289-316). New York: Springer.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10*, 333-344.

van der Linden, W. J. & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika, 75*, 120-139.

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

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

von Davier, M. & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology, 3*, 115-124.

von Davier, M. & von Davier, A. A. (2011). A general model for IRT scale linking and scale transformations. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 225-242). New York: Springer.

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

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H. & Mislevy, R. J. (2000). Item response theory, item calibration and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 61-100). Mahwah, NJ: Lawrence Erlbaum Associates.

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