7 Increasing test efficiency in an international assessment of teachers' general pedagogical knowledge through multidimensional adaptive testing

Andreas Frey[1,2] and Aron Fink[1]

[1]Educational Psychology Faculty, Goethe University Frankfurt, Germany

[2]Centre for Educational Measurement at the University of Oslo

This chapter discusses the potential of multidimensional adaptive testing (MAT) for increasing the measurement efficiency of large-scale assessments. It outlines the building blocks of MAT and describes the configuration of a MAT design for the Teacher Knowledge Survey assessment module, including recommendations for its pilot study, field trial, and main study. A Monte Carlo simulation study is used to illustrate the potential of such a design for the module. The chapter concludes with six concrete recommendations for using MAT to transform the module's knowledge assessment into a very modern, innovative and, at the same time, highly efficient measurement instrument.

Introduction and problem definition

The Teacher Knowledge Survey (TKS) was originally developed by the Centre for Educational Research and Innovation (CERI) in 2015 as a stand-alone survey. As detailed in Chapter 1, the revised assessment framework aims to assess general pedagogical knowledge on three key dimensions and six sub-dimensions:

1. **Instruction**: teaching methods and lesson planning, and classroom management
2. **Learning**: motivational-affective dispositions, and learning and development
3. **Assessment:** evaluation and diagnostic procedures, and data use and research literacy.

More than 200 items were developed to assess this multidimensional structure. Fifty-two of these items were used in the TKS pilot study, which was conducted from April to June 2016 in five countries (Sonmark et al., 2017[1]). These 52 items are dichotomously scored simple multiple choice (MC), i.e. a question with four response options – one correct and three incorrect, and complex multiple choice items (CMC), i.e. a question with four or more response options, each response option has to be answered with "right" or "wrong", "suitable" or "unsuitable". Thirty-three of the items were considered to have appropriate psychometric quality for future use, with each sub-dimension covered by at least three items.

The responses gathered were scaled using item response theory (IRT) (van der Linden, 2016[2]), e.g. with the unidimensional one-parameter logistic model (1PL). The 1PL IRT model describes test items in terms of only one-parameter, item difficulty, *b*, and provides an estimate of the latent ability level *θ* needed for solving the item. Each of the three dimensions was scaled separately. Reliability analyses of the complete pooled sample of lower secondary teachers (*N* = 943) resulted in values for Cronbach's Alpha of 0.55 (instruction), 0.63 (learning), and 0.52 (assessment). These values for the internal consistency of the scales are below the precision thresholds typically deemed appropriate for test score reporting.

Sonmark and colleagues (2017[1]) discussed that the precision of the test results could be increased by using the two-parameter logistic model (2PL), instead of the 1PL. The 2PL describes the probability that an individual with latent ability level *θ* endorses an item with two item characteristic parameters: item difficulty, *b*, and item discrimination *a* (how well an item is able to discriminate between persons differing in ability levels). The authors also noted that it would be useful to obtain more information on the sources of missing data, for example, by tracking the time spent on viewing pages and giving responses. Missing data is problematic, as it reduces statistical power, can reduce the representativeness of the samples, and can cause bias in the estimation of parameters. Improper handling of missing values may lead to inaccurate inference about the data. Finding out the sources of missing data and using appropriate missing data estimation methods is, therefore, of great importance to safeguard the validity of test score interpretations. Another refinement is the revision and extension of the existing item pool, for example, to include polytomous scored items and more situation-based items, in order to assess the more practical aspects of teacher knowledge (see Chapters 1 and 4).

Building on CERI's TKS, the TKS assessment module will form an optional module for the next cycle of the Teaching and Learning International Survey (TALIS) in 2024. The goals for the TKS assessment module are high: While reducing the testing time from 60 (pilot study) to 30 minutes, the reliability, which did not meet common reporting standards in the CERI TKS pilot study, has to be increased substantially. In fact, the proportion of systematic variance in the test scores needs to be roughly doubled. The Cronbach's Alphas, which ranged from 0.52 (systematic variance = $0.52^2 = 0.27$) to 0.63 (systematic variance = $0.63^2 = 0.40$), should be increased to a value of at least 0.75 (systematic variance = $0.75^2 = 0.56$). When such a portion of systematic variance (due to the responses to the cognitive items) is combined with responses to background questionnaire items (and other background variables) in a latent regression approach, a precision adequate for result reporting will be achieved. In order to achieve or approach these ambitious goals, the possibilities of psychometrics and test administration using digital technology (offline and online) must be used in the best possible way; more specifically, by:

1. Using an IRT model that provides higher statistical information (allowing for a more precise measurement of teacher knowledge) while allowing for stable parameter estimates (e.g. 2PL instead of 1PL).
2. Making use of the correlation between the TKS dimensions to increase measurement precision by adopting a multidimensional IRT framework (e.g. multidimensional 2PL [M2PL]).
3. Presenting items with optimised information for each tested teacher by using multidimensional computerised adaptive testing.
4. Making the best out of the available testing time by selecting items that provide maximum statistical information per time unit.
5. Reducing the proportion of time needed to read and process item stimuli to the complete testing time by incorporating units with within-item adaptivity.

While the first two points can be achieved with the current design of the TKS assessment module, the last three points require a multidimensional adaptive testing (MAT) design. The first section of this chapter introduces briefly multidimensional IRT models and describes the key elements of multidimensional adaptive testing (MAT). It explains how MAT can help to cover a broad range of topics within a limited testing time and discusses further advantages and disadvantages of the approach. This section also outlines the six building blocks of adaptive testing that need to be accounted for when planning a MAT design. The next section discusses recommendable usages of MAT for the TKS assessment module. Suggestions are then substantiated with a Monte Carlo simulation study, which was conducted for this expert chapter. After that, the requirements for implementing such a design in terms of software and analytical skills are outlined. The chapter closes by summarising the main conclusions for the TKS assessment module.

What is multidimensional adaptive testing?

How can MAT help to cover a broad range of topics within a limited testing time?

Computerised adaptive testing (CAT) is a special approach to the assessment of latent traits (e.g. teacher knowledge), in which the selection of the test items that are presented next to the test taker is based on the test taker's responses to previously administered items (Frey, 2020[3]). The aim of this selection procedure is to tailor the item presentation to the trait level of the test taker (e.g. teachers' level of general pedagogical knowledge) in order to administer only those items that provide as much diagnostic information as possible about the individual characteristics to be measured.

The main advantage of CAT compared to non-adaptive testing, in a statistical sense, is the possibility of a considerable increase in measurement efficiency (Segall, 2005[4]). This efficiency gain can be used to increase measurement precision if the number of items is held constant for all test takers or it can be used to reduce the test length. Compared to traditional non-adaptive tests, the number of items can typically be reduced by approximately half when CAT is used while comparable measurement precision can be achieved [e.g. Segall (2005[4])]. In addition, CAT provides the possibility to overcome the problem that conventional tests typically measure test takers with average performance much more precisely than low- and high-performers. This is achieved by aligning the standard errors of the ability estimates across the ability range [e.g. Frey and Ehmke (2007[5])].

Adaptive tests can also be used to make the test-taking experience of the participants more positive. For the Programme for International Student Assessment (PISA), for example, a study showed that low-performing students were confronted with a test situation in which they could not solve many of the presented items. This was accompanied by significantly lower levels of test-taking motivation, combined with significantly higher levels of boredom/daydreaming compared to average- and high-performing students (confronted with items with a more appropriate difficulty level). By adaptively adjusting the difficulty level of the items that are presented to the individual test takers, these systematic effects can be avoided (Asseburg and Frey, 2013[6]).

Although unidimensional CAT has proven to be advantageous in many simulation studies and empirical applications, the performance-related constructs (literacies, competencies, abilities, knowledge etc.) conceptualised in international large-scale assessments (ILSAs) are usually quite complex and can seldom be described by a single latent trait. Typically, the theoretical frameworks underlying these constructs include several interrelated components. To reflect this theoretical complexity directly in the measurement procedure, MAT [e.g. Frey and Seitz (2009[7])] can be used. In contrast to unidimensional CAT, with MAT, multiple dimensions can be measured simultaneously and, therefore, a much better fit between the theoretical underpinnings of complex constructs and test content can be achieved within a reasonable testing time. As measurement models, multidimensional item-response-theory (MIRT) [e.g. Reckase (2016[8]); see Box 7.1] models are used in MAT. These models make it possible to include assumptions about the theoretical structure of the construct of interest in the test instrument. Consequently, the resulting test scores can be interpreted clearly with regard to the theoretical framework and differentiated information on multiple dimensions can be reported.

Box 7.1. Multidimensional item-response-theory models

A general MIRT model is the multidimensional three-parameter logistic (M3PL) model, which specifies the probability that an examinee *j* = 1, …, *N* will answer an item *i* correctly ($U_{ij} = 1$) as a function of the ability vector $\boldsymbol{\theta}_j = (\theta_1, \theta_2, \dots, \theta_p)$ for *p* measured dimensions and for item parameters $\mathbf{a}_i', b_i$, and c_i:

$$P\left(U_{ij} = 1 | \boldsymbol{\theta}_j, \mathbf{a}_i', b_i, c_i\right) = c_i + (1 - c_i)\frac{\exp\left(\mathbf{a}_i'(\boldsymbol{\theta}_j - b_i\mathbf{1})\right)}{1+\exp\left(\mathbf{a}_i'(\boldsymbol{\theta}_j - b_i\mathbf{1})\right)}. \quad (1)$$

The loading of item *i* on the different dimensions is represented by the 1 x *p* item discrimination vector $\mathbf{a}_i'$. Depending on whether the items reflect one (between-item multidimensionality) or multiple (within-item multidimensionality) dimensions, one or multiple elements of $\mathbf{a}_i'$ are different from zero. The difficulty of item i is given by the parameter b_i. The pseudo-guessing parameter c_i can be regarded as a lower asymptote that is introduced to model item-specific random guessing.

The multidimensional two-parameter logistic (M2PL) model and the one-parameter logistic (M1PL) model can be derived from the M3PL model shown in Equation 1. The M2PL model is derived from the assumption that, for all test items, c_i is equal to zero. In addition to this, the M1PL model is derived by constraining one or more elements (reflecting between- or within-item multidimensionality) of the vector $\mathbf{a}_i'$ to a non-zero constant and the remaining elements to zero for each item. Besides these standard MIRT models, more complex multidimensional models such as the non-compensatory MIRT model (Hsu and Wang, 2019[9]), the higher-order model (Wang and Kingston, 2019[10]), the multidimensional testlet model (Frey, Seitz and Brandt, 2016[11]), and the scaling individuals and classifying misconceptions model (Bao et al., 2021[12]) can be used for MAT.

The high measurement efficiency of MAT results on the one hand from the same advantages as those of unidimensional CAT mentioned above. On the other hand, additional efficiency gains are achieved by drawing on prior information about the multidimensional distribution of the measured dimensions. This results in an improvement in measurement efficiency, compared to that achieved by using separate unidimensional adaptive tests for each latent dimension [e.g. (Li and Schafer, 2005[13]; Paap, Born and

Braeken, 2019[14]; Segall, 1996[15])]. The measurement efficiency is especially high if latent dimensions are highly correlated [e.g. (Frey, Bernhardt and Born, 2017[16]; Makransky and Glas, 2013[17]; Segall, 1996[15]; Wang and Chen, 2004[18])].

In MAT, both maximum likelihood and Bayesian procedures can be used for the item selection. From these item selection procedures, the Bayesian approach introduced by Segall (1996[15]) has received the most attention in the literature so far. It has also proven to be one of the best performing and very robust methods in terms of accuracy and precision of ability estimates, compared to other item selection methods used across a broad range of MAT configurations (Mulder and van der Linden, 2009[19]; Veldkamp and van der Linden, 2002[20]; Wang and Chang, 2011[21]; Wang, Chang and Boughton, 2011[22]). The Bayesian approach takes into account the fact that individual responses to items constructed to measure one dimension provide information not only about that particular dimension but also about correlated dimensions. In this approach, item selection is optimised by using the variance-covariance matrix $\mathbf{\Phi}$ of the measured latent traits as prior information about the interrelation of the construct's dimensions (e.g. stemming from a field trial). During the test, the candidate item is selected from the item pool that provides the highest increase in measurement precision regarding all dimensions of interest, based on the D-optimality criterion (see Box 7.2 for details).

Box 7.2. D-Optimality

During the test, the item i^* is selected from the item pool that maximises the determinant of the *p* x *p* matrix $\mathbf{W}_{t+i^*}$.

$$|\boldsymbol{W}_{t+i^*}| = \left|\boldsymbol{I}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_j) + \boldsymbol{I}(\boldsymbol{\theta},u_{i^*}) + \boldsymbol{\Phi}^{-1}\right|. (2)$$

$\mathbf{W}_{t+i^*}$ is derived by summing up the information matrix of the previously t administered items $\mathbf{I}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_j)$, the information matrix of a response u_{i^*} to the candidate item i^* $\mathbf{I}(\boldsymbol{\theta},u_{i^*})$, and the inverse of the variance-covariance matrix of the prior distribution of the measured dimensions $\mathbf{\Phi}^{-1}$. To estimate $\widehat{\boldsymbol{\theta}}_j$ during the course of the test, Segall proposes using the multidimensional maximum a posteriori estimator in combination with the same prior information given by $\mathbf{\Phi}$. The candidate item that causes the greatest reduction in the volume of the credibility ellipsoid (multidimensional Bayesian equivalent of a confidence interval) of the current estimated latent ability vector $\widehat{\boldsymbol{\theta}}_j = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p)$ of person *j* regarding the *p* latent dimensions is selected next for administration.

What factors need to be accounted for in planning a MAT design?

As for all adaptive tests, the six building blocks of CAT (Frey, forthcoming[23]) need to be specified: (1) item pool, (2) start of the test, (3) person parameter estimation, (4) item selection, (5) constraint management, and (6) end of the test. All these aspects should be specified based on data from the field trial and pre-operational simulation studies.

Develop a *calibrated multidimensional item pool*

The functionality and quality of a multidimensional adaptive test is highly dependent on the quality of the multidimensional item pool (Building block 1). A clear definition of the content areas to be examined in terms of an assessment framework is a prerequisite for successful item construction and test development. The item pool should be constructed and/or selected (e.g. from existing sets of items) in such a way that it enables an optimal coverage across assumed dimensions and sub-dimensions. In addition, the item pool should allow a broad range of trait levels (e.g. skills) to be targeted for each of the dimensions measured by the test. An adaptive test can adapt optimally if, for every ability level that can occur in the course of the test, there are at least as many items as the test is long.

In order for MAT to work at its maximum performance, the adaptive algorithm needs to update the ability estimation after each item response. In this process, the responses are typically scored automatically. However, that does not mean that human-coded items (e.g. questions with an open response format) cannot be used in MAT. These items can be selected based on the provisional ability vector and the response given can be stored. This response does not provide any additional information for the selection of the next item but can be easily scored by a human coder after the test and used for the final analyses of the response data.

An important decision that needs to be made in the test development phase, and which should also inform item construction, is which MIRT model to use. As in the unidimensional, non-adaptive case, it is important to consider how many item parameters to include in the model (e.g. M1PL or M2PL) and how to treat the item responses (dichotomous/polytomous). As mentioned earlier, when taking multidimensionality into account, a decision has to be made about whether the model allows for between- or within-item multidimensionality.

A good fit with the assessment framework is not the only thing to consider when choosing a suitable MIRT model; it should also be kept in mind that the higher the complexity of the MIRT model, the higher the number of item parameters that need to be estimated and, therefore, the larger the sample size needed for a stable estimation of these parameters. This estimation, also referred to as calibration, is an essential step in MAT development because the resulting item parameter estimates are used as fixed parameters in the operational phase of the adaptive test. For the frequently used 2PL model, for example, a minimum of 500 responses per item is recommended (De Ayala, 2009[24]). As the M2PL can use the correlation between the modelled dimensions in the estimation, this minimum requirement should also be realistic for the M2PL, at least if the number of dimensions is not too high (concrete suggestions for the TKS assessment module are presented below).

Furthermore, an important aspect that has to be considered in planning a MAT design is item position effects (IPEs). IPEs are variations in item parameter estimates that are related to the position in which they are presented in a test. A typical pattern in conventional testing is that the difficulties of the same items tend to increase towards the end of the test.

IPEs can easily be estimated based on traditional ILSA data, which use a balanced booklet design in which the position of items in the booklets is systematically varied. A booklet is the test form which includes the set of items given to the test taker. A lot of empirical evidence underlines that IPEs must be expected in ILSAs, with a typical decreasing proportion of correct answers and, thus, an increase in item difficulties towards the end of a test [e.g. (Albano, 2013[25]; Debeer et al., 2014[26]; Nagy et al., 2019[27]; Wu et al., 2019[28])]. IPEs are even more problematic for adaptive testing than for sequential testing. This is because in adaptive testing, different persons respond to different items in different positions, while the same item parameters are used for all positions for item selection and ability estimation. Ignoring existing IPEs can result in systematic bias in ability estimation and can therefore jeopardise the interpretations of the test results; it is thus a threat to validity [see Frey and Fink (forthcoming[29]) for an in-depth discussion of this problem].

In general, there are two options for dealing with IPEs in adaptive testing: statistical design and statistical modelling. The prerequisite for both options is that all items are presented in all possible positions during the calibration of an adaptive test. It is desirable that items and positions are stochastically independent from each other. One way to realise this is a randomised item selection. However, the desired uniform distribution of the items across positions can only be achieved asymptotically with a very large number of test takers. A second option to achieve the stochastic independence of items and positions is to use a position-balanced booklet design [e.g. Frey, Hartig and Rupp (2009[30])]. Such a booklet design consists of several booklets, with each booklet containing a subset of the item pool. Across all booklets, each item (or each group of items) is presented in each position with equal frequency. By scaling items based on the responses stemming from such a position-balanced booklet design, the resulting difficulties are composed of the individual item difficulty plus the average IPE across positions. During an adaptive test, this leads to biased ability estimates unless the test length of the operational adaptive test equals the test length of the calibration test.

On the basis of these considerations, Frey, Bernhardt and Born (2017[16]) showed how statistical modelling of IPEs could be used to control for unwanted IPEs. They introduced a multistep procedure, which allows the incorporation of parameterised IPEs into the adaptive test if there is empirical evidence for their existence. This flexibility, however, comes with higher sample size requirements for a stable estimation of the IPE parameters and is therefore not optimally suited to ILSAs. An approach that is easier to implement is that introduced by Frey and Fink (forthcoming[29]). Frey and Fink's approach is based on statistical design principles and showed very good performance, including controlling for IPEs, within a Monte Carlo simulation under typical ILSA settings.

Determine specifications for final computerised adaptive testing

Next to a calibrated item pool, the remaining building blocks of an adaptive test need to be specified before its operational use. These building blocks are the start of the test (i.e. which items to select at the beginning of the test), person parameter estimation, item selection, constraint management, and end of the test. This should not be done based on arbitrary decisions; instead, Monte Carlo simulation studies should be conducted. Monte Carlo simulations are essential to compare and evaluate different methods and specifications for the building blocks for a given assessment situation (e.g. available testing time, content to be covered, number of dimensions, etc.) given an already calibrated item pool. Especially different configurations of the item selection algorithm should be simulated at this point.

Besides reaching statistical optimality, the item selection has to take different non-statistical constraints into account. These are, for example, the proportion of items per sub-domain, as well as the item and stimulus type (e.g. picture, video, text), the grouping of several items to units (testlets) that should be kept intact, and much more. The currently most powerful constraint management method for MAT is the shadow-test approach [(Veldkamp and van der Linden, 2002[20]); see Box 7.3 for details]. It enables the simultaneous consideration of a high number of such constraints and provides very good results regarding constraint violations.

Box 7.3. Shadow testing approach

The shadow-test approach (van der Linden and Reese, 1998[31]; Veldkamp and van der Linden, 2002[20]) is based on the idea of selecting items from a hypothetical test (=shadow-test), which is compiled automatically before the selection of each item, instead of selecting from the complete item pool. The algorithm can be described as follows:

1. Initialise the ability estimation.
2. Assemble shadow-test that accounts for all constraints (e.g. test length, content coverage, proportion of items per item type, etc.), contains all already administered items, and is optimal at the current provisional ability estimate.
3. Administer an eligible item from the shadow test.
4. Update ability estimation.
5. Update constraints to consider the non-statistical attributes of the items already administered.

6. Return all unused items to the item pool.
7. Repeat Steps 1–6 until the termination criterion of the adaptive test is met.

As each shadow test at each step is assembled to meet all non-statistical constraints imposed, the resulting set of presented items also meets all constraints. In addition, the shadow-tests are assembled to be optimal regarding the provisional ability estimation at each step. The shadow-test has to be assembled in real time before the administration of each item. This process is handled by automated test assembling methods (van der Linden, 2005[32]) that use a mathematical programming technique called mixed-integer programming. A detailed description of the shadow-test procedure for MAT can be found in (Veldkamp and van der Linden, 2002[20]).

In addition, item exposure constraints should be implemented in the adaptive test. Adaptive item selection that is solely based on statistical optimality will lead to items with high item discrimination parameters being selected very often. Thereby, they have an increased probability to be communicated among potential test takers and to thus become known, which is typically not wanted for items in ILSAs. To avoid this, different exposure control methods integrate a type of randomisation into the item selection [e.g. Huebner et al., (2016[33]); see Box 7.4 for an example of an exposure control method]. In addition to the constraints mentioned above, constraints regarding item response times [e.g. utilising the simplified version of the maximum information per time unit by Cheng, Diao and Behrens, (2017[34])] can be included in the test in order to maximise diagnostic information within a fixed testing time.

Box 7.4. Exposure control with the Sympson-Hetter method

With the Sympson-Hetter method (Sympson and Hetter, 1985[35]), an item is administered to an individual test taker only if it passes a probability experiment. Otherwise, it is removed from the item pool. Therefore, the user specifies a target proportion per item as a parameter for the item selection algorithm. For example, if test developers do not want an item to be administered to more than 50% of the test takers, they specify a probability of 0.5 for each item in the item pool. Each time the algorithm selects an item from the item pool, a random number between 0 and 1 is generated and compared to the specified probability. If the number is between 0 and 0.5, the item is administered. If not, the item is removed from the remaining pool for this particular test taker.

As exposure rates differ substantially between items, there is no need to specify the same probability for each item. Items with a difficulty near 0 or with high discrimination parameters are likely to be presented very often and therefore might deserve a lower probability (e.g. 0.3) than items with extreme item difficulties that are administered only to the top 5% of the test takers. Such items should not be constrained (which equals setting the probability to 1). In order to determine item-specific exposure probabilities with the Sympson-Hetter method, simulation studies are typically carried out.

What could multidimensional adaptive testing designs for the Teacher Knowledge Survey assessment module look like?

Item development

Depending on the resources available for item construction and calibration, an item pool size of 5 to 10 times the test length per dimension is recommended. This means, for example, when administering 10 items per dimension, the complete item pool of well-functioning calibrated items should contain between 150 and 300 items. The more items, the better the test can adapt to the individual trait level. However, the advantages in terms of the measurement precision that can be achieved by increasing the size of the item pool follow a saturation curve. This means that, at the beginning, increasing the item pool size has a large effect, but this effect becomes smaller the more items are added. Even with item pools that are two to three times as large as the test length, considerable gains in measurement precision can be achieved compared to non-adaptive sequential testing [e.g. Spoden, Frey and Bernhardt (2018[36])].

Two possibilities for the development of the item pool seem especially feasible for the TKS assessment module: (1) An item pool consisting of single items only and (2) an item pool consisting of single items plus sets of items connected to a shared innovative stimulus, for example video or text vignettes of typical classroom situations (see Chapter 4 for a discussion of assessments using vignettes), or interactive stimuli, with adaptive item selection within units. From a statistical point of view, a multidimensional adaptive test with single items (possibility 1) has optimal flexibility to adapt. However, given the relatively short testing time available for the TKS assessment module (30 minutes), the items need to have relatively short stimuli in order for the testing time to be used efficiently. This might be problematic because measuring some aspects of teacher knowledge [e.g. knowledge-based decision making in the classroom and teachers' classroom management expertise; (Stürmer and Seidel, 2015[37]; König, 2015[38])] is likely to require more complex stimulus material. Innovative single items with video or text vignettes as stimuli, for example, could be a remedy here but would be too time consuming and may jeopardise reaching an appropriate level of measurement precision. Therefore, possibility (2) represents an innovative alternative, which also meets the wish to include more complex situation-based items, for example, a few video or text vignettes of typical classroom situations across countries and economies. For these, a larger number of items (e.g. 25) covering a broad difficulty range and different sub-dimensions of the assessment framework could be developed. These items can be regarded as a unit-specific item pool. During the adaptive test, the stimulus of such an innovative unit is presented and items (e.g. eight) are selected from the unit-specific item pool according to an item selection criterion. Each test taker is presented with one innovative unit and the rest of the testing time is filled with adaptively selected single items. If several innovative units are constructed that cover the assessment framework well, a good content coverage will be achieved across test takers. For all items that are constructed anew, the typical item development procedures, including cognitive labs, should be carried out.

Specification of the multidimensional adaptive testing design

As the psychometric model, the three-dimensional 2PL model (or the generalised partial credit model, GPCM, in the case of polytomous items) with between-item multidimensionality could be used to measure the three broad dimensions of general pedagogical knowledge (instruction, learning and assessment) specified in the TKS assessment framework. This model provides considerably higher statistical information than the three-dimensional 1PL but minimises the potential problem of item parameter estimates varying between countries/economies or assessments that is more likely to occur when more complex models are used. Item selection based on D-optimality and using maximum a posteriori (MAP) estimation of the provisional ability during the adaptive test is recommended. In order to maximise the statistical information obtained in the given testing time, item selection criteria that take the response times of the individual test takers into account could also be considered. A viable representative of such an item selection criterion, whose performance, however, has not yet been examined in the context of MAT in ILSAs, is the simplified maximum information per time unit criterion suggested by Cheng et al. (2017[34]).

In order to reflect the assessment framework within each adaptive testing session, the content constraints that have to be taken into account are the three dimensions of the TKS (instruction, learning and assessment), each of which is composed of two additional sub-dimensions. The main dimensions should each be measured by an equal number of items. To reach a content coverage that conforms with the assessment framework, within each dimension, the sub-dimensions should also be equally represented. In addition, the three transversal aspects (knowledge about using technology, fostering 21st century skills and managing diversity in classrooms) should be equally represented across dimensions. These content constraints could be considered simultaneously and automatically by using the shadow-test approach (see Box 7.3). Exposure control methods, such as the Sympson-Hetter method (see Box 7.4), should be used to avoid over- and underexposure of some items.

The starting items could be chosen randomly from a set of items with difficulties near zero, or if applicable, teacher responses to background questionnaires could be used to determine the starting point of the test (e.g. if teachers indicate that they are novice teachers, they would get an easier starting block of items than their experienced colleagues). Termination criteria should be a test length of 30 items or a testing time of 30 minutes; whatever is reached first. Both test length and testing time should be kept constant between the field trial and the main study in order to be able to control for IPEs. In order to fine-tune the adaptive algorithm, pre-operational Monte Carlo simulations are recommended prior to the main study, based on the empirical results from the field trial.

After all responses had been gathered, the final scaling should be conducted. For this scaling, the M2PL is recommended in conjunction with a latent regression approach and drawing of plausible values (PVs), as done in PISA (OECD, forthcoming[39]). The PVs form the basis for the calculation of the reported results. The next three sections cover recommendations for the pilot study, the field trial and the main study of the TKS assessment module.

Pilot Study

As stated above, it would be useful to expand the item pool of the TKS assessment module, which currently consists of over 200 items. Especially units with innovative situation-based stimuli (e.g. video vignettes) with several (25 or more) connected items covering a broad difficulty range would make the TKS assessment module a very future-oriented and modern assessment. These items can be developed using standard item development procedures as they are typically used for OECD large-scale assessments. It is mandatory to pilot these innovative units. In addition, it would be useful to also include the existing TKS items in the pilot study. The IRT scaling of the gathered responses should be conducted with the M2PL. On the basis of the scaling results, deficient items can be identified and improved, if possible, or excluded from the item pool. The resulting variance-covariance matrix (before conditioning with a background model) and the provisional item parameter estimates will then be used in the field trial. The aim would be to have an item pool of about 150 good, calibrated items that covers all components of the assessment framework.

Field Trial

As the number of available items is too large for them all to be presented to one test taker, a balanced incomplete block design (BIBD) could be used to assemble different test versions. BIBDs are the type of design that was used, for example, for the paper-based assessments of PISA up to 2012. For the case of the TKS assessment module, the design can be similar to the design used by Spoden, Frey and Bernhardt (2018[36]) in the construction process of a three-dimensional computerised adaptive test. This design has two levels: At the first level, a Youden square design [e.g. Giesbrecht and Gumpertz, (2004[40])] is used for each dimension (here: instruction, learning and assessment). So, for each dimension, it is ensured that across all test versions (1) the test length is kept constant, (2) all items are presented with equal frequency, (3) each pair of items is presented with equal frequency, and (4) each item is presented in every possible position with equal frequency to control for IPEs. This first level design is nested in the second level.

At the second level, three blocks, one for each dimension, are specified. Therefore, for the case of the TKS, each test version would comprise one block of items for instruction, one block of items for learning, and one block of items for assessment. This balanced block design balances potential order effects at the second level. Thereby, it is ensured that (5) each test contains one block for each dimension and (6) each possible ordering of the three blocks is used with equal frequency across test versions. The composition of the individual test versions can easily be done by computer. Note that a design of comparable quality can only be approximated by manually generated, labour-intensive multistage testing designs.

Defining sample size requirements is typically not trivial because they depend on a multitude of conditions. In order to make this chapter as concrete as possible, a conservative proposal with regard to the sample size is formulated below. However, it will certainly also be possible to achieve good results with other - possibly smaller - sample sizes. For a stable estimation of the M2PL model, there should be 500 or more responses per item, per country/economy. For an item pool with 150 items and a test length of 30 items, this would require a calibration sample size of at least *N* = 2,500 test takers. By using online calibration designs, such as the balanced continuous calibration strategy [CCS; (Fink et al., 2018[41]; Frey and Fink, forthcoming[29]), see below], this sample size requirement can be reduced, but it should not fall below 100 responses per item, per country/economy. In the example with an item pool of 150 items, this would lead to a sample size requirement of at least *N* = 500 teachers in the field trial per country/economy. By using such a calibration strategy, the field trial can be used to get an initial set of item parameter estimates. Clearly deficient items can be identified via item fit and differential item functioning (DIF) analyses across countries/economies and can then be excluded from the main study. The item parameters and the latent variance-covariance matrix estimated from the field trial can be used for multidimensional adaptive item selection and provisional ability estimation during the adaptive test administration in the main study.

Main Study

The same test system, test length, and testing time as in the field trial should be used in the main study. The use of the CCS is recommended, with proportions of items per dimension and per sub-dimension controlled for by shadow testing. The CCS includes concurrent scaling using the responses to all non-drifted items (items with substantial differences in item parameter estimates across assessments) from previous assessments (here: field trial) while controlling for IPEs, and it leads to a fast and continuous improvement of the item parameter estimates. The CCS therefore provides a good compromise between a stable estimation of item parameters and an optimisation of measurement precision. The resulting data can be used for the typical psychometric analyses such as those of item fit, country/economy DIF, and others. Using the final item parameter estimates, the results can be estimated and reporting that is based on PVs obtained by scaling with a latent regression approach is possible. Using the CCS makes it easy to add items to future assessments if needed and to improve item parameter estimates on the fly while controlling for IPEs.

Simulation of efficiency and precision gains for the Teacher Knowledge Survey assessment module

In order to obtain an impression of the efficiency gains that can be expected from using the suggested MAT design, a simulation study was conducted. For this purpose, a simplified version of the suggested MAT design was compared to two non-adaptive test designs: one based on a unidimensional IRT model (*urand*) and the other one on a multidimensional IRT model (*mrand*). The simulation study assumed an overall item pool of 150 items (50 items per dimension) under the M2PL model. Item difficulties were generated by extending the difficulty parameters obtained from the pilot study, as reported in Sonmark et al. (2017[1]), by drawing randomly from a standard normal distribution, $b \sim N(0,1)$. Discrimination parameters were randomly drawn from a lognormal distribution, with $a \sim logN(0,.25)$. Each item loaded on exactly one dimension (between-item multidimensionality). For the field trial, the study simulated *N* = 500 test takers (simulees). The ability parameters of the simulees were randomly drawn from a multivariate normal distribution using a conservative estimate of the mutual correlation between the TKS dimensions of 0.70:

$$\boldsymbol{\theta} \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Phi}), \text{ with } \boldsymbol{\mu} = (0,0,0) \text{ and } \boldsymbol{\Phi} = \begin{bmatrix} 1 & 0.70 & 0.70 \\ 0.70 & 1 & 0.70 \\ 0.70 & 0.70 & 1 \end{bmatrix}.$$

The responses were generated for a linked calibration design, with the items assigned to 10 subsets of 15 items (five for each dimension). Each form consisted of two of these subsets, with one common subset between Forms 1 and 2, Forms 2 and 3, and so on. Forms were administered in a balanced way, in order to obtain 100 responses per item during calibration. Item parameters were estimated using marginal maximum likelihood (MML; Bock and Aitkin, 1981[42]). These item parameter estimates were used for adaptive item selection in the MAT condition.

After calibration, the adaptive and the two non-adaptive tests were simulated. Responses were simulated for a main study sample of *N* = 2,000 simulees. True abilities were randomly drawn from a multivariate normal distribution:

$$\boldsymbol{\theta} \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Phi}), \text{ with } \boldsymbol{\mu} = (0,0,0) \text{ and } \boldsymbol{\Phi} = \begin{bmatrix} 1 & 0.70 & 0.70 \\ 0.70 & 1 & 0.70 \\ 0.70 & 0.70 & 1 \end{bmatrix}.$$

The test length for each condition was set to 30 items, with 10 items per dimension. For non-adaptive testing, items were randomly drawn from the item pool. For MAT, the D-optimality criterion was used for item selection. Information regarding the correlation between the three dimensions was incorporated into the adaptive item selection and ability estimation (for the MAT and mrand conditions) by using the variance-covariance-matrix $\boldsymbol{\Phi}$ estimated from the calibration data. MAP (Mislevy, 1986[42]) was used for ability estimation. In order to obtain more uniform item exposure rates across the complete item pool, the Sympson-Hetter exposure control method was integrated into the adaptive item selection. Afterwards, response matrices gathered from the simulated field trial and the main study in each condition were combined and the final item parameters were estimated using MML estimation. On the basis of these item parameters, final ability parameters were estimated using MAP estimation. For each condition, r = 20 replications were compared regarding the resulting test information (overall and per dimension), averaged across simulees given their true ability levels. In addition, reliability, calculated as the squared correlation between true and estimated ability (Kim, 2012[43]), was calculated for each dimension. The simulation was carried out in R (R Core Team, 2020) using the package mirtCAT (Chalmers, 2016[44]) to simulate the tests and the package mirt (Chalmers, 2012[45])for item and person parameter estimation. Table 7.1 shows the results of the simulation.

Table 7.1. Test information and reliability per simulation condition averaged across replications

Test information			Reliability		
Instruction	Learning	Assessment	Instruction	Learning	Assessment
urand					
3.931	4.014	3.804	0.653	0.639	0.660
(0.017)	(0.015)	(0.014)	(0.008)	(0.008)	(0.008)
mrand					
3.931	4.014	3.804	0.721	0.706	0.720
(0.017)	(0.015)	(0.014)	(0.010)	(0.009)	(0.008)
MAT					
4.667	4.764	4.463	0.753	0.747	0.763
(0.416)	(0.372)	(0.423)	(0.020)	(0.022)	(0.015)

Note: urand = random item selection and unidimensional IRT model; mrand = random item selection and multidimensional IRT model; MAT = multidimensional adaptive testing; standard errors are given in parentheses.

As the urand and the mrand condition used the same response matrix, the resulting test information with regard to the true ability was the same in these conditions. It can be seen that MAT provided a substantial increase in test information. The effect of using MIRT modelling instead of unidimensional IRT modelling is reflected in the increase in reliability between the urand and the mrand conditions. The additional effect of multidimensional adaptive item selection and the associated higher test information is illustrated by the increase in reliability from the mrand to the MAT condition. The results demonstrate that, even with a relatively small item pool of five times the test length, a short test length, and the integration of multiple constraints into the adaptive item selection, MAT increases the measurement precision up to a range that is well suited for precise result reporting, while this is not the case for non-adaptive testing. Nevertheless, it has to be noted that the simulation study did not include missing responses, which have to be expected when applying the TKS assessment module to real teachers. Therefore, it can be assumed that the test information and the reliability will be somewhat lower for empirical data, while the relative differences between the conditions are likely to be the same.

What is required in terms of software and analytical skills for implementing such designs?

All methods and algorithms needed to implement a highly efficient MAT design are already implemented in packages in the statistical programming language R (R Core Team, 2020[46]). It would thus be straightforward to build an adaptive TKS assessment module based on R. If an existing testing platform should be used for item delivery, an interface between this platform and R would need to be programmed. As R is free and open-source, it enables the integration of self-programmed algorithms as well as the usage and adaptation of already existing algorithms, according to the needs of test developers. For example, the mirt package (Chalmers, 2012[45]) can be used to conduct MIRT analyses including scaling, fit analysis, DIF analysis, latent regression analysis using plausible values (PVs), and others. In addition, the mirtCAT package (Chalmers, 2016[44]) can be used for adaptive item selection. The mirtCAT package includes a large variety of item selection methods as well as the possibility to customise the adaptive algorithm in accordance with the test-specific requirements. Furthermore, modern constraint management and exposure control methods, such as the shadow-test approach and the Sympson-Hetter method, can be applied using the package. Besides these two packages, some sub-routines of the KAT-HS-App (Fink et al., forthcoming[47]), which is also programmed in R, can be integrated into the test system to impose the CCS as described in (Frey and Fink, forthcoming[29]).

Implementing the suggested combination of methods requires psychometric expertise in IRT modelling and CAT development. The former comprises technical skills that are not that different from the skills that are usually required for traditional, non-adaptive ILSAs (e.g. IRT scaling and linking, fit analysis, item analysis, DIF analysis). The latter requires, alongside technical skills in IRT, comprehensive knowledge about calibration designs, adaptive algorithms (including item selection criteria, constraint management, ability estimation), skills in R-programming, and experience in conducting Monte Carlo simulations. However, because all suggested methods are already implemented in frequently used R packages, no specialised programming expertise is needed.

Conclusion

The aim of this chapter was to discuss possibilities to increase the measurement efficiency of the TKS assessment module for future cycles of TALIS by using state-of-the-art psychometric approaches and computer-based test administration in a goal-oriented way. In order to substantially increase the reliability while reducing the testing time of the TKS pilot study by 50%, six points are suggested to achieve this ambitious goal. The first two points can be achieved with the current test design, whereas the other four require a change to a MAT design (which, thus, might only be implemented in later cycles of the module):

1. The 2PL model should be used instead of the 1PL model. The 2PL provides considerably higher statistical information than the 1PL while still allowing for a stable estimation of item parameters.
2. The M2PL model, as the multidimensional extension of the 2PL, should be used to further increase the measurement precision by using information about the correlation between the three dimensions covered by the TKS assessment module. The results of the initial simulation study presented in this chapter show that, even with a conservative estimate of a mutual correlation of 0.70 between the TKS dimensions and even when using random item selection, a substantial gain in reliability can be achieved when using the M2PL instead of the 2PL model.
3. MAT should be used in order to administer only highly informative items to each individual test taker. The results of the initial simulation showed that using MAT with as few as 10 items per dimension results in a precision level that is appropriate for reporting, even with a correlation between dimensions of only 0.70. It is expected that the latent correlation between the three dimensions of the TKS will be even higher. This would lead to a further increase in measurement efficiency, which can be used, for example, to place more constraints on the test content or to compensate for not-simulated factors such as missing responses.
4. Teachers' responses to background questionnaires should be used to determine the starting point of the test. More precisely, this means that novice teachers would get a different starting block of items than experienced teachers.
5. It would be worthwhile to consider an item selection procedure that takes individual response times into account in order to maximise information per time unit. Such a procedure was not covered in the simulation presented here but is likely to result in further small improvements in terms of measurement efficiency.
6. The incorporation of innovative units with scenario-based stimuli and within-item adaptivity is recommended. This will not only make the assessment modern and future-oriented but will also lead to a very efficient usage of the longer processing time needed for more complex, situation-based items.

All methods needed have already been developed and published and can be applied with free statistical software. As shown, it would already be possible to use TALIS 2024 as a starting point for the adaptive TKS assessment module and to continue it with future cycles, even if some participating countries and economies change (see Table 8.1 in Chapter 8 for the main takeaways from this chapter for TALIS and the TKS assessment module). Contractors whose staff have solid psychometric training will be able to perform the analyses needed to implement and operate the suggested MAT design. Even though the implementation of the six proposed points requires some additional effort, this chapter and the results from the simulation study should encourage test developers to use computers in the best possible way, in order to create an innovative and psychometrically optimised assessment.

References

Albano, A. (2013), "Multilevel modeling of item position effects", *Journal of Educational Measurement*, Vol. 50/4, pp. 408-426, http://dx.doi.org/10.1111/jedm.12026. [25]

Asseburg, R. and A. Frey (2013), "Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit", *Psychological Test and Assessment Modeling*, Vol. 55/1, pp. 92–104. [6]

Bao, Y. et al. (2021), "Flexible computerized adaptive tests to detect misconceptions and estimate ability simultaneously", *Applied Psychological Measurement*, Vol. 45/1, pp. 3–21, http://dx.doi.org/10.1177/0146621620965730. [12]

Chalmers, R. (2016), "Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications", *Journal of Statistical Software*, Vol. 71/5, pp. 1-39, http://dx.doi.org/10.18637/jss.v071.i05. [44]

Chalmers, R. (2012), "Mirt: a multidimensional item response theory package for the R environment", *Journal of Statistical Software*, Vol. 48/6, pp. 1-29, http://dx.doi.org/10.18637/jss.v048.i06. [45]

Cheng, Y., Q. Diao and J. Behrens (2017), "A simplified version of the maximum information per time unit method in computerized adaptive testing", *Behavior Research Methods*, Vol. 49, pp. 502-512, http://dx.doi.org/10.3758/s13428-016-0712-6. [34]

De Ayala, R. (2009), *The Theory and Practice of Item Response Theory*, Guilford, New York. [24]

Debeer, D. et al. (2014), "Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment", *Journal of Educational an Behavioral Statistics*, Vol. 39/6, pp. 502-523, http://dx.doi.org/10.3102/1076998614558485. [26]

Fink, A. et al. (2018), "A continuous calibration strategy for computerized adaptive testing", *Psychological Test and Assessment Modeling*, Vol. 60/3, pp. 327–346. [41]

Fink, A. et al. (forthcoming), *Kriteriumsorientiertes adaptives Testen mit der KAT-HS-App [Criterion-referenced adaptive testing using the KAT-HS-App]*. [47]

Frey, A. (2020), "Computerisiertes adaptives Testen [Computerized adaptive testing]", *in Testtheorie und Fragebogenkonstruktion*, Springer, Berlin, Heidelberg, http://dx.doi.org/10.1007/978-3-662-61532. [3]

Frey, A. (forthcoming), *Computerized adaptive testing and multistage testing*. [23]

Frey, A., R. Bernhardt and S. Born (2017), "Umgang mit Itempositionseffekten bei der Entwicklungcomputerisierter adaptiver Tests [Handling of item position effects in the development of computerized adaptive tests]", *Diagnostica*, Vol. 63, pp. 167-178, http://dx.doi.org/10.1026/0012-1924/a000173. [16]

Frey, A. and T. Ehmke (2007), "Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards [Hypothetical Implementation of Adaptive Testing for the Assessment of Educational Standards]", *Zeitschrift für Erziehungswissenschaft*, Vol. 8, pp. 169–184, http://dx.doi.org/10.1007/978-3-531-90865-6_10. [5]

Frey, A. and A. Fink (forthcoming), *Controlling for item position effects when adaptive testing is used in large-scale assessments*. [29]

Frey, A., J. Hartig and A. Rupp (2009), "An NCME Instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice", *Educational Measurement: Issues and Practice*, Vol. 28/3, pp. 39–53, http://dx.doi.org/10.1111/j.1745-3992.2009.00154.x. [30]

Frey, A. and N. Seitz (2009), “Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges”, *Studies in Educational Evaluation*, Vol. 35, pp. 89–94, http://dx.doi.org/10.1016/j.stueduc.2009.10.007. [7]

Frey, A., N. Seitz and S. Brandt (2016), “Testlet-based multidimensional adaptive testing”, *Frontiers in Psychology*, Vol. 18/1758, pp. 1-14, http://dx.doi.org/10.3389/fpsyg.2016.01758. [11]

Giesbrecht, F. and M. Gumpertz (2004), *Planning, Construction, and Statistical Analysis of Comparative Experiments*, John Wiley & Sons, Inc., Hoboken, NJ. [40]

Hsu, C. and W. Wang (2019), “Multidimensional computerized adaptive testing using non-compensatory item response theory models”, *Applied Psychological Measurement*, Vol. 43/6, pp. 464-480, http://dx.doi.org/10.1177/0146621618800280. [9]

Huebner, A. et al. (2016), “Item exposure control for multidimensional computer adaptive testing under maximum likelihood and expected a posteriori estimation”, *Behavior Research Methods*, Vol. 48, pp. 1443–1453, http://dx.doi.org/10.3758/s13428-015-0659-z. [33]

Kim, S. (2012), “A note on the reliability coefficients for item response model-based ability estimates”, *Psychometrika*, Vol. 77/1, pp. 153–162, http://dx.doi.org/0.1007/S11336-011-9238-0. [43]

König, J. (2015), “Measuring classroom management expertise (CME) of teachers: A video-based assessment approach and statistical results”, *Cogent Education*, Vol. 2/1, pp. 1-15, http://dx.doi.org/10.1080/2331186X.2014.991178. [38]

Li, Y. and W. Schafer (2005), “Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics”, *Applied Psychological Measurement*, Vol. 29/1, pp. 3–25, http://dx.doi.org/10.1177%2F0146621604270667. [13]

Makransky, G. and C. Glas (2013), “The applicability of multidimensional computerized adaptive testing for cognitive ability measurement in organizational assessment”, *International Journal of Testing*, Vol. 13/2, pp. 123–139, http://dx.doi.org/10.1080/15305058.201. [17]

Mislevy, R. (1986), “Bayes modal estimation in item response models”, *Psychometrika*, Vol. 51/2, pp. 177–195, http://dx.doi.org/10.1007/BF02293979. [42]

Mulder, J. and W. van der Linden (2009), “Multidimensional adaptive testing with optimal design criteria for item selection”, *Psychometrika*, Vol. 74/2, pp. 273–296, http://dx.doi.org/10.1007/S11336-008-9097-5. [19]

Nagy, G. et al. (2019), “A multilevel study of position effects in PISA achievement tests: Student- and school-level predictors in the German tracked school system”, *Assessment in Education: Principles, Policy & Practice*, Vol. 26/4, pp. 422–443, http://dx.doi.org/10.1080/0969594X.2018.1449100. [27]

OECD (forthcoming), *PISA 2018 Technical Report*, OECD Publishing, Paris. [39]

Paap, M., S. Born and J. Braeken (2019), “Measurement efficiency for fixed-precision multidimensional computerized adaptive tests: Comparing health measurement and educational testing using example banks”, *Applied Psychological Measurement*, Vol. 43/1, pp. 68-83, http://dx.doi.org/10.1177/0146621618765719. [14]

R Core Team (2020), *R: A language and environment for statistical computing [Software]*, Vienna: R Foundation for Statistical Computing, https://www.R-project.org/. [46]

Reckase, M. (2016), “Logistic multidimensional models”, *in Handbook of Item Response Theory*, Chapman & Hall/CRC, Boca Raton. [8]

Segall, D. (2005), “Computerized adaptive testing”, *in Encyclopedia of Social Measurement*, Elsevier, Amsterdam. [4]

Segall, D. (1996), “Multidimensional adaptive testing”, *Psychometrika*, Vol. 61/2, pp. 331–354, http://dx.doi.org/10.1007/BF02294343. [15]

Sonmark, K. et al. (2017), “Understanding teachers’ pedagogical knowledge: report on an international pilot study”*, OECD Education Working Papers*, No. 159, OECD Publishing, Paris, https://dx.doi.org/10.1787/43332ebd-en. [1]

Spoden, C., A. Frey and R. Bernhardt (2018), “Implementing three CATs within eighteen months”, *Journal of Computerized Adaptive Testing*, Vol. 6/3, pp. 38–55, http://dx.doi.org/10.7333/1809-060338. [36]

Stürmer, K. and T. Seidel (2015), “Assessing professional vision in teacher candidates: Approaches to validating the observer extended research tool”, *Zeitschrift für Psychologie*, Vol. 223/1, pp. 54–63, http://dx.doi.org/10.1027/2151-2604/a000200. [37]

Sympson, J. and R. Hetter (1985), “Controlling item-exposure rates in computerized adaptive testing”, *in Controlling Item-Exposure Rates in Computerized Adaptive Testing*, Navy Personnel Research and Development Center, San Diego. [35]

van der Linden, W. (2016), *Handbook of Item Response Theory*, Chapman & Hall/CRC, Boca Raton. [2]

van der Linden, W. (2005), *Linear Models for Optimal Test Design*, Springer, New York, NY. [32]

van der Linden, W. and L. Reese (1998), “A model for optimal constrained adaptive testing”, *Applied Psychological Measurement*, Vol. 22/3, pp. 259-270, http://dx.doi.org/10.1177/01466216980223006. [31]

Veldkamp, B. and W. van der Linden (2002), “Multidimensional adaptive testing with constraints on test content”, *Psychometrika*, Vol. 67/4, pp. 575–588, http://dx.doi.org/10.1007/BF02295132. [20]

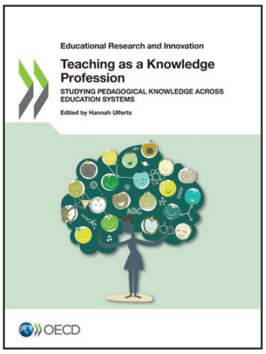
Wang, C. and H. Chang (2011), “Item selection in multidimensional computerized adaptive testing—Gaining information from different angles”, *Psychometrika*, Vol. 76/3, pp. 363–384, http://dx.doi.org/10.1007/S11336-011-9215-7. [21]

Wang, C., H. Chang and K. Boughton (2011), “Kullback-Leibler Information and its applications in multidimensional adaptive testing”, *Psychometrika*, Vol. 76/1, pp. 13–39, http://dx.doi.org/10.1007/s11336-010-9186-0. [22]

Wang, W. and P. Chen (2004), “Implementation and measurement efficiency of multidimensional computerized adaptive testing”, *Applied Psychological Measurement*, Vol. 28/5, pp. 295–316, http://dx.doi.org/10.1177/0146621604265938. [18]

Wang, W. and N. Kingston (2019), “Adaptive testing with a hierarchical item response theory model”, *Applied Psychological Measurement*, Vol. 43/1, pp. 51-67, http://dx.doi.org/10.1177/0146621618765714. [10]

Wu, Q. et al. (2019), “Predictors of individual performance changes related to item positions in PISA assessments”, *Large-scale Assessments in Education*, Vol. 7/5, pp. 1-21, http://dx.doi.org/10.1186/s40536-019-0073-6. [28]


Educational Research and Innovation
Teaching as a Knowledge Profession
STUDYING PEDAGOGICAL KNOWLEDGE ACROSS EDUCATION SYSTEMS
Edited by Hannah Ulferts
OECD

From:

Teaching as a Knowledge Profession

Studying Pedagogical Knowledge across Education Systems

Access the complete publication at:
https://doi.org/10.1787/e823ef6e-en

Please cite this chapter as:

Ulferts, Hannah (ed.) (2021), “Increasing test efficiency in an international assessment of teachers’ general pedagogical knowledge through multidimensional adaptive testing”, in *Teaching as a Knowledge Profession: Studying Pedagogical Knowledge across Education Systems*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/f6d2f568-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.