Psychological Test and Assessment Modeling, Volume 55, 2013 (1), 81-91

Effect of item order on item calibration and item bank construction for computer adaptive tests

Otto B. Walter[1] *& Matthias Rose*[2]

Abstract

Item banks are typically constructed from responses to items that are presented in one fixed order; therefore, order effects between subsequent items may violate the independence assumption. We investigated the effect of item order on item bank construction, item calibration, and ability estimation. 15 polytomous items similar to items used in a pilot version of a computer adaptive test for anxiety (Walter et al., 2005; Walter et al., 2007) were presented in one fixed order or in a order randomly generated for each respondent. A total of n=520 out-patients participated in the study. Item calibration (Generalized Partial Credit Model) yielded only small differences of slope and location parameters. Simulated test runs using either the full item bank or an adaptive algorithm produced very similar ability estimates (expected a posteriori estimation). These results indicate that item order had little impact on item calibration and ability estimation for this item set.

Key words: item response theory; computer adaptive testing; local independence; item bank construction

[1] *Correspondence concerning this article should be addressed to:* Otto B. Walter, PhD, Universität Bielefeld, Fakultät für Psychologie und Sportwissenschaft, AE Psychologische Methodenlehre und Qualitätssicherung, Postfach 100131, 33501 Bielefeld, Germany; email: otto.walter@charite.de

[2] Charité Universitätsmedizin Berlin, Medizinische Klinik mit Schwerpunkt Psychosomatik, Berlin, Germany

1. Introduction

Local item independence is a central assumption of almost any application of Item Response Theory models. Items are locally independent if for respondents at the same level of the underlying latent trait θ responses to any given item are independent of responses to other items of the test (Henning, 1989). Local independence does not prevent items from correlating across the range of all observed ability levels, but it does imply lack of correlation among items if the ability level is fixed. Therefore, local independence is a way to state that it is indeed the latent trait that explains the relations between item responses. Local independence may be violated if other person parameters such as other latent traits are involved in the responses. If this is the case, responses have to be explained by multiple latent variables rather than by one underlying latent trait only, and, therefore, the application of a unidimensional item response model may no longer be appropriate. Lack of independence can also ensue if the response to one item is no longer independent of the responses to previous items. This type of response dependence can occur when previous items contain clues to following items and item order obviously plays an important role here. In the literature, these two types of item dependence, trait multidimensionality and response dependence, are often not clearly distinguished from each other and checking an item bank for local independence is often simply referred to as "ensuring unidimensionality". This is particularly true when unidimensional item response models are used, which, despite the rising interest in multidimensional item response models (e.g. Reckase, 2009), are still dominant in practical applications of item response theory such as the construction of item banks for computer adaptive testing. Table 1 shows the steps required to construct an item bank for unidimensional computer adaptive testing (Walter, 2010). Local item independence and item order play a crucial role in this process. For the construction of the item bank, the order of presentation of the items is typically fixed. In an adaptive test, the item selection algorithm determines the order of presentation and this order can vary for each respondent.

The purpose of the present study is to investigate the impact of item order on item bank construction. The general idea is to compare item parameter estimates obtained from responses given to items presented in fixed order with item parameters estimated from responses given to items that were presented in random order. Numerical differences in item parameter estimates may or may not have significant impact on ability estimates. Practitioners are usually much more interested in ability levels of respondents rather than in item parameters. The focus of this study is, therefore, on quantifying how much ability level estimations differ when item banks are constructed from responses to items given in fixed versus in random order. To assess this effect, a simulation study was conducted using the two item banks obtained from item presentation in fixed and random order. In simulated adaptive tests, estimates of person parameters for fictitious respondents with known ability levels (simulees) were compared for these two item banks.

Table 1:
Development steps of a unidimensional computer-adaptive test (adapted from Walter, 2010).

Present block of items to participants in fixed order
Record and score responses
Ensure local item independence
Investigate item response curves
Check items for DIF
Calibrate items
Present items in order determined by CAT process

2. Method

2.1 Sample

The development of the item banks derived from items presented in fixed and random order was based on data obtained from *n*=520 out-patients of the Department of Psychosomatic Medicine, Charité, University Medicine Berlin, Germany between July 2004 and January 2005. In addition to the psychometric assessment in the department, these patients answered 15 items with five response options similar to items of a computer adaptive test for anxiety (Walter et al., 2005; Walter et al., 2007). Data collected during these assessments were used to conduct empirical item analyses and the item bank calibration described below. Data from follow-up assessments were not used in this study. About two thirds of the patients were female (female: 65.4%, male: 34.6%). The mean age was 41.1 years (*SD*: 12.5 years). According to the main clinical ICD-10 diagnoses, the sample was comprised of 21% depressive disorders (F32-34/F45), 19% somatoform disorders (F45), 13% anxiety disorders (F40/F41), 10% eating disorders and addictions (F10/F50/F55), and 9% somatic diseases. The rest of the patients (28%) suffered from other conditions but were neither assigned to a main ICD-10 diagnosis in the F group nor a main somatic diagnosis.

2.2 Data collection

All items were administered in a computer-assisted way using personal digital assistants (PDAs). These palm-sized devices were equipped with a touch screen on which the items were presented separately in German language (Rose et al., 2002). One half of the PDAs was prepared to present the 15 items in fixed order, the other half of the PDAs was set up to generate a new random permutation of the 15 items for each assessment. The two sets of PDAs were indistinguishable from each other; the ward secretary had no knowledge which set a PDA belonged to and randomly chose a PDA to hand over to the patient.

2.3 Item analyses

The item banks for both the fixed and the random item order group were constructed similarly to the steps we used to construct computer adaptive tests for anxiety, depression and stress; a detailed description can be found in Walter et al. (2005, 2007) (Anxiety-CAT), Fliege et al. (2005) (Depression-CAT), and Kocalevent et al. (2009) (Stress-CAT). To ensure unidimensionality, we conducted a one-factorial confirmatory factor analysis for categorical variables using a standardized solution and the weighted least square means and variance adjusted (WLSMV) estimator (Mplus version 3.1, Muthén & Muthén, 2004), and excluded one item of each pair of items exhibiting a correlation in the residual correlation matrix larger than 0.20 (Bjorner, Kosinski & Ware, 2003). Item response curves computed non-parametrically were inspected visually (Gaussian kernel smoothing; Ramsay, 1995). This step aims at comparing the shapes of observed item response curves with those of parametrically modeled functions. Ideally, a category function should exhibit steep trace lines with one sharp maximum that exceeds all other response functions in exactly one interval of the latent trait. Sorted in ascending order, the θ values for which a response function is maximal should match the order in which the response choices are presented.

2.4 Item calibration and item banks

The 15 items used in this study were revised versions of items contained in an item bank of a computer adaptive test for anxiety (Walter et al., 2007, 2009). The original item bank was constructed from various German questionnaires or versions of international questionnaires in German language indicative of anxiety. As the items were drawn from a diverse pool of instruments, several response formats were present in the items. The revision of the items aimed at harmonizing these different response formats across the items in the pool but some differences were still present after the revision. In the study presented here, only items with five ordered response categories were considered.

The items of the fixed and random order group were calibrated separately using the Generalized Partial Credit Model (GPCM; Muraki, 1992), a two-parameter model for polytomous items. Item parameter estimation was conducted by the marginal maximum likelihood procedure implemented in the PARSCALE software (Muraki & Bock, 1999). In this model, the probability of endorsing response k ($k \in \{0,\ldots,K\}$) of item j with $K+1$ response options is given by

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{i=1}^{k} \alpha_j(\theta - \beta_{ji})\right]}{\sum_{m=0}^{K} \exp\left[\sum_{r=0}^{m} \alpha_j(\theta - \beta_{jr})\right]}$$

The parameter θ denotes the level of the latent trait, α_j is the slope parameter of item j, and $\beta_{j0},\ldots,\beta_{jK}$ stand for the threshold parameters of this item. The first threshold parameter is set zero ($\beta_{j0} = 0$).

The item calibration yielded two item banks: item bank A comprised the item parameters of the items presented in fixed order and item bank B contained the item parameters estimated from the same items presented in random order.

2.5 Simulation study and adaptive algorithm

To quantify the amount to which the two item banks yield different person parameter estimates, a simulation study was conducted. For each 0.2 interval of the latent trait between -3.0 and +3.0 responses of n=100 simulees were generated using the method described by Wang (1999; details are also provided in Walter et al., 2007). For both item banks, person parameter estimates were computed using the adaptive algorithm described below.

The adaptive test algorithm used in this study consists of the following steps. (1) In the first step, the person parameter estimate is set to zero, which is the assumed population mean. (2) For the current estimate of the person parameter, the item with the highest Fisher-Information is selected and presented to the respondent. (3) The response to this item is used to compute a new estimate of the person parameter and standard error using the expected a posteriori (EAP) method with a standard normally distributed prior (Bock & Mislevy, 1982). Steps (2) and (3) are repeated until either the current standard error falls below a given threshold (0.32) or all items in the item bank have been presented to the respondent.

Each simulation was run with two stopping rules: (a) no stopping rule (i.e. use of all items in the item banks), and (b) standard error (*SE*) of the person parameter estimate smaller than 0.32 (corresponding to a reliability greater than 0.9).

3. Results

3.1 Item analyses

The fixed order and the random order group comprised n=239 and n=281 respondents respectively. One-factorial confirmatory factor analyses for categorical variables yielded low residual correlations below 0.20 for the majority of item pairs in both groups with the exception of two items that showed residual correlation above this threshold in both groups. These two items pertaining to relaxation (“I am at ease” and “I am relaxed”) were excluded from further analysis. Subsequent confirmatory factor analyses in both groups showed no residual correlations above 0.20 for the remaining items.

Visual analysis of item response curves computed non-parametrically yielded very similar trace lines for the majority of corresponding items. An ideal item exhibits steep trace lines and sharp maxima. As a rule of thumb, the general steepness of the trace lines corresponds to the slope parameter of the GPCM (the steeper the trace lines, the higher the slope parameter); the positions of the intersections of the trace lines of adjacent response

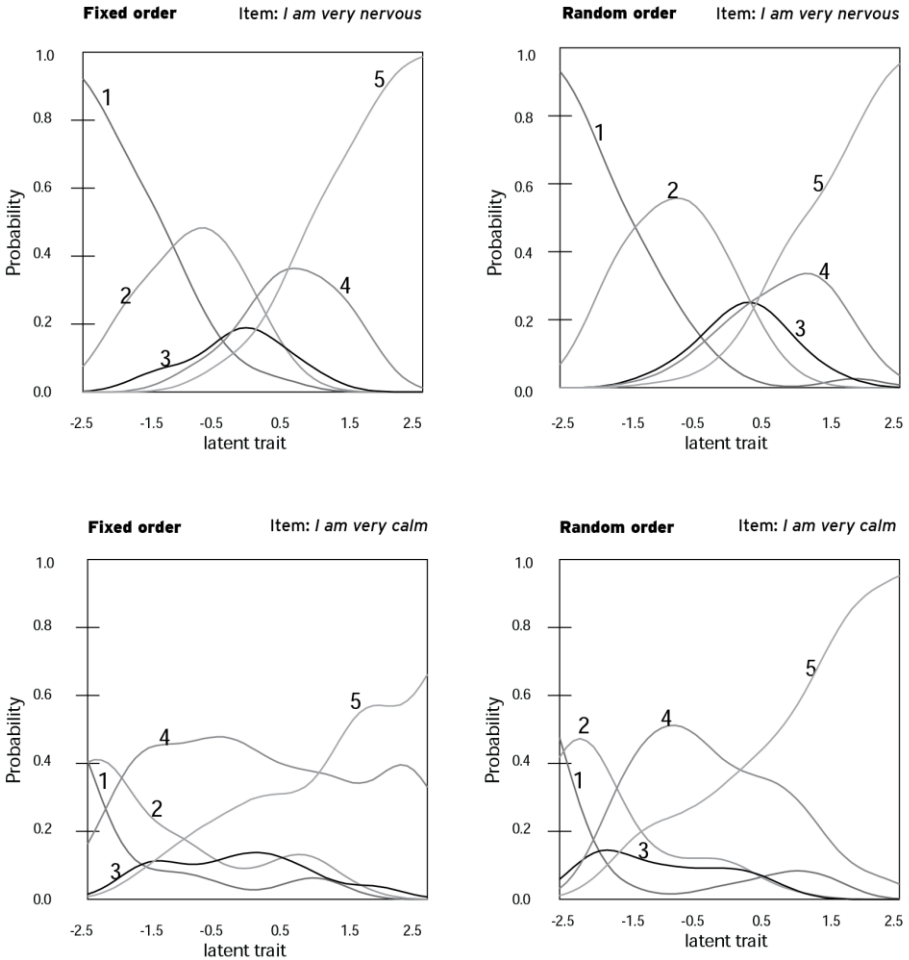

Figure 1:

Analysis of two items using Gaussian kernel smoothing presented in fixed (left) and random order (right). The five trace lines of each item correspond to the probability of choosing one of the five response categories as a function of the latent trait.

Top: The item in position 6 when presented in fixed order (response categories: 1 *not at all*, 2 *somewhat*, 3 *more or less*, 4 *very much so*, 5 *exactly*). The item exhibits similar trace line patterns (steep curves and peaked maxima) for both fixed order (top left) and random order (top right) presentation.

Bottom: The item that was presented first in fixed order (response categories: 1 *exactly*, 2 *very much so*, 3 *more or less*, 4 *somewhat*, 5 *not at all*). In comparison to the item presented at the top of the figure, the trace lines of this item are less steep and do not show peaked maxima. However, presentation of this item at the beginning of the test (bottom left) aggravates this pattern considerably and yields an even stronger deviation from the ideal trace line pattern than presentation in random order (bottom right).

categories on the latent trait can serve as ballpark figures of the location parameters of the GPCM. Fig. 1 (top) shows an example item. This item was presented in position 6 when presented in fixed order and similar trace lines were found for both fixed and random order presentation. Applying the rule of thumb, item calibration for fixed and random order presentation should yield values of the slope and location parameters that are similar (see next section). The only two items for which the item response curves differed more noticeably were the two items that were presented first and second in the fixed order group. Fig. 1 (bottom) illustrates this effect for one of those two items, namely the item that was presented first in the fixed order group. This item appears to have trace lines that are not as steep as those of the item shown at the Fig. 1 (top). However, the deviation from the ideal pattern is much more pronounced when this item is presented at the beginning of the test.

3.2 Item calibration

A two-parameter model for polytomous items was employed to calibrate the items (Generalized Partial Credit Model [GPCM]). The metric was set in reference to a population mean of 0 and a standard deviation of 1. The 13 items in both groups had five response categories. Therefore, item calibration using the GPCM yields one slope parameter and four threshold parameters for each item. The slope parameter determines, to a great extent, the item information, which, in turn, is decisive with regard to selecting an item according to the maximum information rule and governs overall measurement precision. Both the slope parameter and the threshold parameters estimates were very similar for corresponding items of the two groups. For instance, the slope parameter estimates of the item that was presented in position 6 in fixed order (cf. Fig. 1, top) were 1.40 for fixed and 1.35 for random order of presentation. Overall, the mean difference between slope parameters in the fixed and random order group was -0.04 (*SD*: 0.49). The only two items for which the item parameters estimates showed more pronounced differences were the two items that were presented as first and second items in the fixed order group. For instance, the slope parameter estimates for the item that was presented first in fixed order (cf. Fig 1, bottom) were 0.29 for fixed and 0.38 for random order of presentation.

3.3 Simulation study

To assess the impact of the (mostly small) numeric differences between the item parameters in the two banks, a simulation study was conducted. Latent trait estimates (EAP estimation) obtained from both administering all 13 items and employing an adaptive algorithm were computed for the two item banks. Fig. 2 (top) shows the estimated values from all available items in the pools as a function of the true level of latent trait. This relation is S-shaped because of the bias towards the prior mean of the EAP estimation (Chen, Hou & Dodd, 1998; Meijer & Nering, 1999). However, this bias is noticeable only for extreme values of the latent trait θ (about $|\theta| > 2$), and becomes negligible for less extreme values (Bock & Mislevy, 1982). More importantly, though, both item banks yielded very similar estimates in the whole θ range, indicating that the small differences

in item parameters had little impact on the estimation of person parameters. This is also true for a simulated adaptive algorithm (Fig. 2, bottom). Again, the person parameter estimates between -2 and 2 Logits were very similar for both item banks. For $|\theta| \leq 2$, where about 95% of a population under the standard normal distribution is expected to score, both item banks required only 7.9 items (*SD*: 3.4) to estimate the person parameter

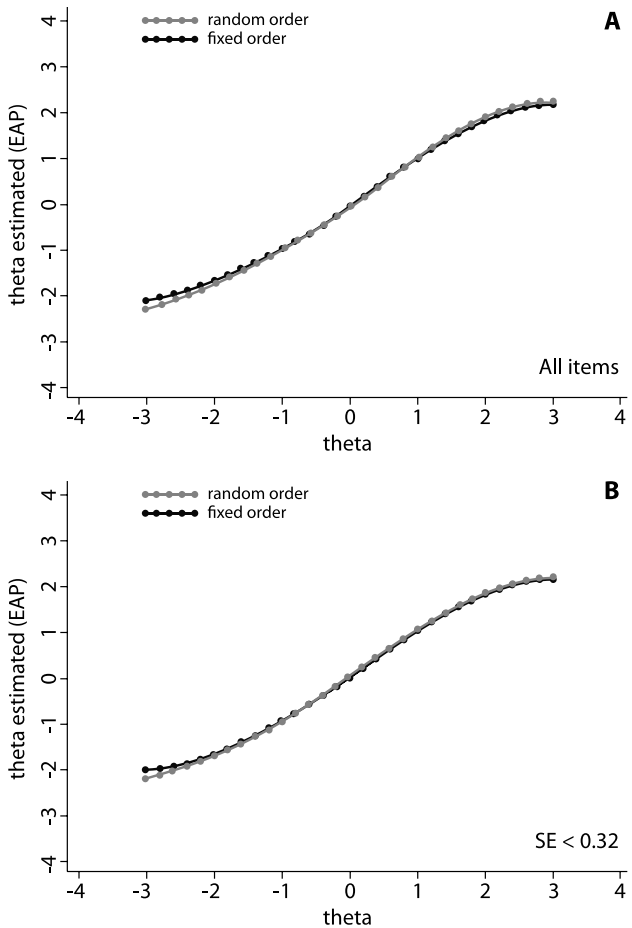

Figure 2:
Plot of the simulated latent trait (θ) estimates (EAP estimation) against true abilities of simulees for the item banks constructed from fixed and random order presentation. Each dot on the y-axis is the mean of *n*=100 estimates of simulees with the latent trait level shown on the x-axis.

Top (A): Trait levels estimated from all available items in the item banks.

Bottom (B): Trait levels estimated using a computer adaptive process and a stopping rule set to *SE* < 0.32.

with high precision. For this range of the latent trait, about 40 % of the items were saved because of the adaptive algorithm. For more extreme values of θ, item savings due to the adaptive algorithm were considerably less pronounced. In the whole range of latent trait between -3 and +3, differences of person parameter estimates computed from the two item banks are hardly noticeable even when an adaptive algorithm is employed.

4. Discussion

The study presented here aimed at investigating the effect of item order on item bank construction. Items were presented in fixed and in random order and were calibrated. For the majority of items, only small differences in item parameters were found and these small differences had hardly any effect on person parameter estimates. Noticeable differences in item response functions and, consequently, also in item parameters were found only for the two items that were presented as first and second items in the fixed order group. This is indicative of a "warming-up effect" in the sense that respondents needed about two items to familiarize themselves with the item format and the item response categories. Two recommendations for item calibration can be drawn from this observation. First, warm-up items may be beneficial to calibrating items in fixed order. These items should mirror the general character of the items to calibrate but should not be used for the actual calibration. For this item set, two warm-up items are needed. Further research is needed to identify conditions that affect the number of warm-up items required. Secondly, administration of items in random order is actually an alternative worth considering when calibrating small item pools in a computer-assisted way. If large item pools have to be calibrated the use of balanced incomplete design would be another possibility (e.g. Yousfi & Böhme, 2012; Frey, Hartig & Rupp, 2009). Visual analysis of item response curves computed using a non-parametric method indicated that the peaks of the response curves tended to have sharper maxima and resembled the shape of parametric item response models more closely. A possible explanation for this effect is that item order effects may cancel out each other when items are presented in changing random permutations, and, therefore, may facilitate the parametric estimation of an item response model. Cautionary notes on item order effects in adaptive testing as raised by Ortner (2008, 2004) should still be taken seriously though. Ortner reported that respondents who were presented with an item representing a high level of the latent trait at the beginning of the test exhibit a tendency to agree with fewer items than those respondents confronted with other items at the beginning of the test (Ortner, 2008). These results underline the need for further research on context and carry-over effects in adaptive testing. Asseburg (2011) tried to understand the processes underlying the response behavior in adaptive testing and her study can be seen as a promising albeit rare example of how such theory-based research may be conducted. Even though our results suggest that item order appears to have less impact on item calibration than one might assume, item order effects can also occur *after* item calibration has taken place. These effects pertain to the reaction of the respondent to the adaptive test situation and should be a preferred focus of further research. Adaptive testing has the promise of improving psychometric

assessment, but the investigation of the psychology of adaptive testing has just begun to identify both the potential and the problems ahead.

5. References

Asseburg, R. (2011). Leistungsbereitschaft in Testsituationen: Motivation zur Bearbeitung adaptiver und nicht-adaptiver Leistungstests. Marburg: Tectum. (Online: http://eldiss.uni-kiel.de/macau/receive/dissertation_diss_00006627).

Bjorner, J., Kosinski, M., & Ware, J. E. (2003). Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT-super™). *Quality of Life Research*, *12*, 913-933.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Chen, S., Hou, L., & Dodd, B. (1998). A comparison of maximum likelihood estimated and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement*, *58*, 569-595.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, *14*, 2277-2291.

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.

Henning, G. (1989). Meanings and implications of local independence. *Language Testing*, *6*, 95-108.

Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, B. J., Kleiber, D., & Klapp, B. F. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology*, *62*, 278-287.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing. Overview and introduction. *Applied Psychological Measurement*, *23*, 187-194.

Muraki, E. (1992). A Generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muraki, E., & Bock, R. D. (1999). *PARSCALE: Analysis of graded responses and ratings*. Chicago, IL: Scientific Software Int., Inc.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus. The comprehensive modeling program for applied rearchers. User's guide*. Los Angeles: Muthén & Muthén.

Ortner, T. (2004). On changing the position of items in personality questionnaires. Analysing effects of item sequence using IRT. *Psychology Science*, *46*, 466-476.

Ortner, T. (2008). Effects of changed item order: A cautionary note to practitioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection and Assessment*, *16*, 249-257.

Ramsay, J. O. (1995). *TestGraf. A program for the graphical analysis of multiple choice test and questionnaire data*. Montreal: McGill University.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Dordrecht: Springer.

Rose, M., Walter, O. B., Fliege, H., Becker, J., Hess, V., & Klapp, B. F. (2002). 7 years of experience using Personal Digital Assistants (PDA) for psychometric diagnostics in 6000 inpatients and polyclinic patients. In H. B. Bludau & A. Koop (Eds.), *Mobile computing in medicine. Lecture notes in Informatics* (pp. 35-44). Bonn: Köllen.

Walter, O. B. (2010). Adaptive Tests for Measuring Anxiety and Depression. In W. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 123-136). Berlin: Springer.

Walter, O. B., Becker J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for "Anxiety" (Anxiety-CAT). *Quality of Life Research*, *16*, 143-155.

Walter, O. B., Becker, J., Fliege, H., Bjorner, J. B., Kosinski, M., Walter, M., Klapp, B. F. & Rose, M. (2005). Entwicklungsschritte für einen computeradaptiven Test zur Erfassung von Angst (A-CAT). *Diagnostica*, *51*, 88-100.

Wang, S. (1999). *The accuracy of ability estimation methods for computerized adaptive testing using the generalized partial credit model* (Unpublished doctoral dissertation). University of Pittsburgh, PA.

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